The Open University

M248

Analysing data

Computer Book C

This publication forms part of an Open University module. Details of this and other Open University modules can be obtained from Student Recruitment, The Open University, PO Box 197, Milton Keynes MK7 6BJ, United Kingdom (tel. +44 (0)300 303 5303; email general-enquiries@open.ac.uk).

Alternatively, you may visit the Open University website at www.open.ac.uk where you can learn more about the wide range of modules and packs offered at all levels by The Open University.

The Open University, Walton Hall, Milton Keynes, MK7 6AA.

First published 2017.

Edited, designed and typeset by The Open University, using the Open University TEX System.

Printed in the United Kingdom by Hobbs the Printers Limited, Brunel Road, Totton, Hampshire SO40 3WX.

ISBN 978 1 4730 2264 5

5.1

Contents

Introduction

This computer book covers all the computer work associated with Book C of M248 *Analysing data*. Most of the computer work involves using Minitab, but there is also one animation to explore.

Using this book

As you study each unit in Book C, you will be directed to work through particular chapters in this book as part of your work on that unit. Each unit contains instructions as to when you should first refer to particular material in this computer book; you are advised not to work on the activities here until you have reached the appropriate points in the units.

As with Computer Books A and B, the activities vary in nature and length. You should try to work through all the activities as you read the chapters; you will find solutions to the activities at the end of the book.

A few supplementary exercises on the whole of this computer book are provided after Chapter 5. You may use these for extra practice or for revision (or not at all), as you wish.

1 The method of least squares

This chapter is associated with Subsection 2.1 of Unit 11.

In this chapter, you will use one of the M248 animations to investigate the least squares line for linear regression data.

Houses built using the method of least squares?

Activity 1 *The method of least squares*

- Open the **Method of least squares** animation.

(a) On the right-hand side of the animation you will see axes for a (currently empty) scatterplot. You can enter a point on the scatterplot by clicking on the scatterplot at the place where you would like the point.

- Try entering a few points on the scatterplot now.

The associated values of x and y for each of the points on the scatterplot are stored (and can be changed) in the table on the left-hand side of the animation. You can edit the values of x or y of any point by selecting the value in the table and editing it, then clicking elsewhere in the table or pressing either the **Enter** key or the **Tab** key.

- Try editing one of your x or y values in the table now and observe how the position of the point changes on the scatterplot.

A point can be moved by dragging it on the scatterplot using the mouse. A point can also be deleted by hovering the mouse pointer over the point and pressing either the **Delete** or **Backspace** key.

- Try moving, and then deleting, a few points on the scatterplot now.

Clicking on the **Reset** button will clear the scatterplot.

- Click on the **Reset** button now.

You can also delete a point by clicking on the row in the table that contains its coordinates and pressing the **Delete** key. Note that if a cell in the table is selected for editing, then you need to press the **Enter** key before you can delete a point on the plot using the **Backspace** key.

(b) In this part of the activity, you will enter a set of points on the scatterplot and then fit a line to these points by eye.

- Enter a set of points on the scatterplot which are scattered roughly about a straight line going from the lower-left corner to the upper-right corner of the scatterplot.

Below the table on the left-hand side of the animation are two rows of buttons. The first row (labelled **Fitted line**) is for fitting lines to the points in the scatterplot. By default, **None** is selected, so initially there are no lines on the scatterplot.

- Click on the **Yours** button.

A horizontal black line with a drag handle at either end will appear on the scatterplot. You can move the line by dragging the drag handles.

- By dragging the drag handles, move the line until (in your opinion) it seems to fit the data well.

Recall that the least squares line minimises the sum of the squares of the residuals.

The second row of buttons (labelled **Residuals**) is for displaying the residuals (by selecting **Lines**), or the squared residuals (by selecting **Squares**), for the chosen line. (You can remove the residuals from the scatterplot by selecting **None**.)

- Click on **Lines** so that you can see the residuals as vertical lines from each of the points to your fitted line.

When either type of residuals is displayed, the residual sum of squares is given below the scatterplot.

- Adjust the position of your line until the residual sum of squares seems to be minimised.

(c) When the **Least squares** button in the top row of buttons is selected, the least squares line is fitted to the points on the scatterplot and is added to the plot using a green line. The residuals to the least squares line are then shown by selecting either **Lines** or **Squares** in the second row of buttons.

- Obtain the least squares line for your set of points.

Notice that when the least squares line is added to the scatterplot, the line you fitted using **Yours** changes to a broken line. How does the least squares line compare with the line that you fitted using **Yours**?

(d) Clear the scatterplot and repeat parts (b) and (c) for several different sets of points to get a feel for the method of least squares.

In the next activity, you will again use the M248 animation **Method of least squares**, this time to explore how the presence of outliers affects the least squares line.

Activity 2 *Outliers*

If necessary, click on the **Reset** button in the animation **Method of least squares** to clear the scatterplot and table.

(a)
- Enter a set of points on the scatterplot that lie roughly on a straight line going from the lower-left corner to the upper-right corner of the diagram.
- Add the least squares line to the scatterplot.
- Now add an extra point in the upper-left corner (or in the lower-right corner) of the diagram.

What happens to the least squares line?

- Delete the extra point and then add a point near the top (or near the bottom) of the scatterplot, but halfway across the diagram.

Describe the effect on the least squares line of adding a point in this position.

(b)
- Clear the scatterplot and enter a new selection of points, all lying in a cluster on the left-hand side of the scatterplot.
- Add one point on the right-hand side of the scatterplot and obtain the least squares line.
- Move the outlying point up and then down.

How does this affect the position of the least squares line?

In Activity 2, you have seen that some outliers influence the least squares line considerably. An outlier that changes the position of the least squares line substantially when it is added to a scatterplot is called an **influential point**.

2 Fitting a linear regression model

This chapter is associated with Subsection 3.2 of Unit 11.

In this chapter, you will use Minitab to fit the least squares line to data and to check whether the assumptions of a linear regression model are reasonable – that is, to check that the random terms are normally distributed with constant, zero mean and constant variance. This is done using **Fit Regression Model...** from the **Regression** submenu of the **Regression** submenu of the **Stat** menu.

Yes, we do mean 'the **Regression** submenu of the **Regression** submenu ...'.

Start up Minitab now if it is not already running.

Activity 3 *Cholesterol data*

Data on the total cholesterol levels measured for 11 individuals aged over 40 years were explored in Unit 11. The least squares line for the cholesterol data was estimated on the basis of summary statistics in Example 11 of that unit, and the assumptions of the model were checked in Examples 13 and 14. Here, you will use Minitab to do both. The data are in the worksheet **cholesterol.mtw**. Open this worksheet now.

- Choose **Stat** > **Regression** > **Regression** > **Fit Regression Model...**. The **Regression** dialogue box will open.
- The response variable is the total cholesterol, so enter the variable `Cholesterol` in the **Responses** field.
- An explanatory variable is called a *predictor* in Minitab. Enter the explanatory variable `Age` in the **Continuous predictors** field.
- Leave the **Categorical predictors** field blank: you won't need to use this field in M248.

A residual plot and a normal probability plot of the residuals for the model you wish to fit can be obtained via the button in the **Regression** dialogue box labelled **Graphs...**.

- Click on **Graphs...** to obtain the **Regression: Graphs** dialogue box.
- Select the options **Normal probability plot of residuals** and **Residuals versus fits** (by clicking on the corresponding tick boxes). You should leave the other options unselected, and ensure that the **Residuals for plots** field at the top of the dialogue box is set to **Regular** (the Minitab default).
- Click on **OK** in the **Regression: Graphs** dialogue box to return to the **Regression** dialogue box, then click on **OK** in the **Regression** dialogue box.

The following output is produced in the Session window.

Regression Analysis: Cholesterol versus Age

```
Analysis of Variance

Source         DF  Adj SS   Adj MS  F-Value  P-Value
Regression      1  0.5843  0.58433     5.52    0.043
  Age           1  0.5843  0.58433     5.52    0.043
Error           9  0.9520  0.10578
  Lack-of-Fit   7  0.8270  0.11815     1.89    0.389
  Pure Error    2  0.1250  0.06250
Total          10  1.5364


Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0.325240  38.03%     31.15%      13.41%
```

```
Coefficients

Term         Coef  SE Coef  T-Value  P-Value   VIF
Constant    1.889    0.911     2.07    0.068
Age        0.0407   0.0173     2.35    0.043  1.00


Regression Equation

Cholesterol = 1.889 + 0.0407 Age


Fits and Diagnostics for Unusual Observations

Obs  Cholesterol    Fit   Resid  Std Resid
  5        3.300  3.926  -0.626      -2.03  R

R  Large residual
```

Despite the large amount of information given above, you need only pick out the equation of the fitted least squares line which is given under '`Regression Equation`':

Just one other part of this Minitab output will also be considered below.

$$\text{Cholesterol} = 1.889 + 0.0407\,\text{Age}.$$

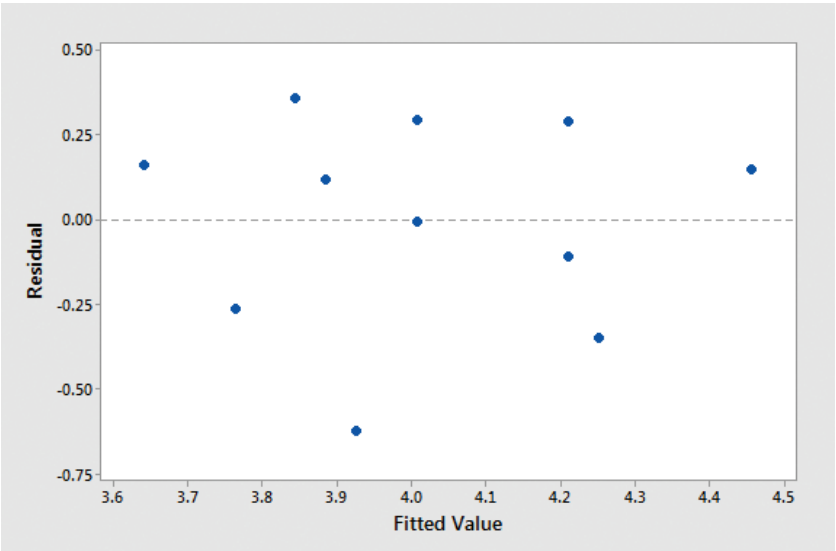
Notice that Minitab has presented the result to a slightly greater accuracy than was done in Unit 11.

The two plots requested in the **Regression: Graphs** dialogue box are reproduced in Figure 1 (overleaf).

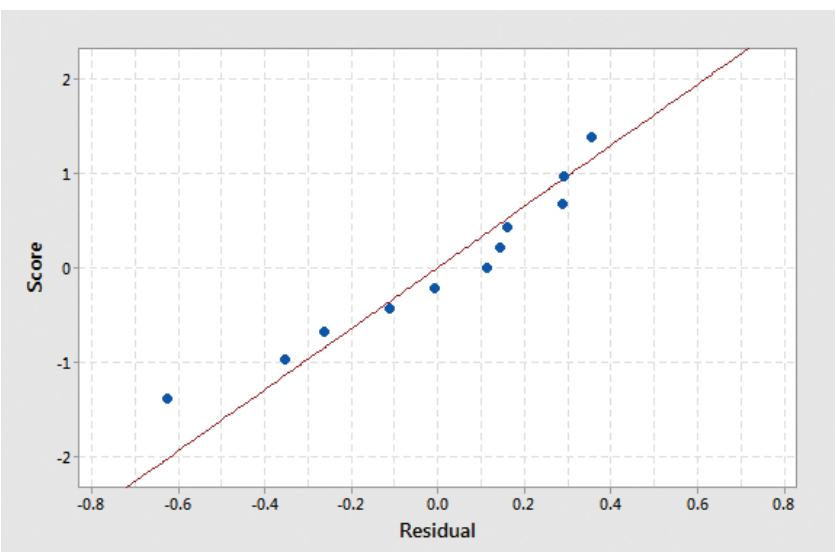
The residual plot in Figure 1(a) is very similar to the residual plot given in Figure 22 of Unit 11. There is no obvious pattern in the plot, thus the assumption that the random terms come from distributions with constant, zero mean and constant variance seems reasonable. We say this despite the final part of the Minitab output in the Session window, which is headed '`Fits and Diagnostics for Unusual Observations`'. This part of the output draws attention to particular points that Minitab considers to be unusual. In this case, it highlights a single point, the one with fitted value 3.926 and residual -0.626, as having a large residual. This is the point that is bottom-most in Figure 1(a). This point does not seem to be sufficiently unusual to be of real concern.

The normal probability plot of the residuals is shown in Figure 1(b). A normal probability plot produced by Minitab as part of a regression analysis uses the method that is currently set as the default method for obtaining the plot points. In Chapter 2 of Computer Book B, the Herd–Johnson method was set as the default method, so the plotted points in Figure 1(b) are the same as those in Figure 26 of Unit 11. The points lie roughly along a straight line so the conclusion is that a normal distribution seems a plausible model for the random terms, although this conclusion is arguable especially around the middle-right. (The large, negative, residual highlighted by Minitab is the point to the bottom-left of the normal probability plot that is also rather out of line with the normality assumption.)

To check and/or change the default, use **Tools** > **Options...** as described in Activity 7 of Computer Book B.



(a)



(b)

Figure 1 (a) Residual plot (b) normal probability plot of residuals

Activity 4 *Racing times and wind speed*

Hitchcock, S.E. (1993) 'Does wind speed affect times?', *Track Stats*, vol. 31, pp. 19–26.

In a study of the effect of wind speed on athletes' race times, the speed of the following wind (in metres per second) and race time (in seconds) were recorded for each of 21 races in 1990 for the British 110 m hurdler Colin Jackson. The data are given in the worksheet **jackson.mtw**. Open the worksheet now. (Negative speeds of the following wind correspond to speeds of a headwind into which the athlete is running.)

(a) Obtain a scatterplot of the race times (in the column `Race time`) against the wind speeds (in the column `Wind speed`), with the race times on the vertical axis. Do you think a linear regression model might be a good model for these data?

(b) Calculate the least squares line for the relationship between race time and wind speed.

(c) Interpret what the values of the least squares estimates of the parameters of the regression line tell us.

(d) Check the assumptions of the fitted model. Are the assumptions reasonable for these data?

Colin Jackson in action

Activity 5 *Road distances*

In this activity, you will fit a linear regression model to the data from Example 8 in Unit 11 on road and map distances (both in miles) between locations in and around Sheffield. In Unit 11, it was suggested that a model for these data should be a straight-line regression model with the constraint that the line goes through the origin. The worksheet **distance.mtw** contains the road and map distances. Open the worksheet now.

In Minitab, the procedure for calculating the least squares line with the constraint that the line must go through the origin is almost the same as for calculating the unconstrained least squares line.

- Choose **Stat** > **Regression** > **Regression** > **Fit Regression Model...** to open the **Regression** dialogue box.
- In the **Regression** dialogue box, enter `Road` in the **Responses** field, and `Map` in the **Continuous predictors** field.
- In the **Regression: Graphs** dialogue box, make sure that **Normal probability plot of residuals** and **Residuals versus fits** are ticked. Check that **Residuals for plots** is set to **Regular**.
- Click on **OK** to close the **Regression: Graphs** dialogue box and to return to the **Regression** dialogue box.

To fit a line through the origin, you need to specify that in the model.

- From the **Regression** dialogue box, click on **Model...** to open the **Regression: Model** dialogue box.

There is a tick box labelled **Include the constant term in the model** at the bottom of the **Regression: Model** dialogue box. When this box is ticked (which is the default), Minitab calculates the equation of the unconstrained least squares line

$$y = \alpha + \beta x.$$

When the box is not ticked, Minitab calculates the least squares line through the origin

$$y = \gamma x.$$

- A line through the origin is required here, so untick **Include the constant term in the model**.
- Click on **OK** to close the **Regression: Model** dialogue box, and then click on **OK** to close the **Regression** dialogue box.

The last part of the output produced in the Session window by Minitab contains the equation of the line that it has fitted:

$$\text{Road} = 1.2891\,\text{Map}.$$

This is the result that was obtained in Example 10 of Unit 11, given correct to one more decimal place.

Now check the assumptions of the fitted model in the same way as you are used to doing for the unconstrained linear regression model. Are the assumptions reasonable for these data?

3 Fitting a multiple regression model

This chapter is associated with Subsection 5.4 of Unit 11.

In this chapter, you will use Minitab to fit a multiple linear regression model and to check whether the assumptions of the model – that is, that the random terms are normally distributed with constant, zero mean and constant variance – are reasonable. As you will discover in the next computer activity, using Minitab for multiple regression is almost identical to using Minitab for linear regression with one explanatory variable, as discussed in Chapter 2.

Activity 6 *Student satisfaction for unaffiliated universities*

Example 16 in Unit 11 introduced a dataset regarding official statistics for 24 UK universities known collectively as Russell Group universities. There are other groups of UK universities, and in this computer activity we'll consider data for UK universities which are unaffiliated to any particular university group: there are 47 such universities. The data for these universities are in the Minitab worksheet **unaffiliated.mtw**. Open this worksheet now.

The Minitab worksheet contains data in five columns. The first column, `University`, contains the name of each university with the data for that university in the associated row. The second column contains data for the response variable student satisfaction (`Satisfaction`). Columns 3–5 contain data for three explanatory variables. The first two explanatory variables are those considered in Example 16: student–staff ratio (`Ratio`) and academic services spend (`Academic spend`). The third explanatory variable, which wasn't considered in Example 16, is the facilities spend (`Facilities spend`). The data for this variable were collected by the

Higher Education Statistics Agency and use the average expenditure over the academic financial years (2012/13, 2013/14 and 2014/15) to allow for uneven patterns of expenditure. Facilities spend was calculated as being the average expenditure, in pounds, on student facilities (sports, careers services, health, counselling, etc.), divided by the number of full-time equivalent students in the latest academic year.

A sport growing in popularity in many universities is quidditch, which has its roots in the Harry Potter fictional game. An important piece of equipment for the sport is a broomstick which each player is supposed to hold between their legs.

Fit a multiple regression model to these data as follows.

- Choose **Stat** > **Regression** > **Regression** > **Fit regression model...**. The **Regression** dialogue box will open.
- The response variable is `Satisfaction`, so enter this in the **Responses** field.
- When fitting a linear regression model with one explanatory variable, the explanatory variable was entered in the **Continuous predictors** field. For multiple regression, all of the explanatory variables can be entered into this same field. So enter the three explanatory variables, `Ratio`, `Academic spend` and `Facilities spend` in the **Continuous predictors** field. (If typing the variable names, you will need to use quotes for `Academic spend` and `Facilities spend` so that Minitab knows that these are single variables.) Leave the **Categorical predictors** field empty.
- Producing a residual plot and a normal probability plot of the residuals for the model is done in exactly the same way as for linear regression, namely, click on **Graphs...** to obtain the **Regression: Graphs** dialogue box and select **Normal probability plot of residuals** and **Residuals versus fits**. Check that **Residuals for plots** is set to **Regular**.
- Click on **OK** in the **Regression: Graphs** dialogue box.
- Click on **Model...** in the **Regression** dialogue box, and ensure that the **Include the constant term in the model** option in the **Regression: Model** dialogue box is ticked.
- Click on **OK** in the **Regression: Model** dialogue box and then click on **OK** in the **Regression** dialogue box.

The following output is produced in the Session window.

Regression Analysis: Satisfaction versus Ratio, Academic spend, Facilities spend

`Analysis of Variance`

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	0.150955	0.050318	5.65	0.002
Ratio	1	0.050909	0.050909	5.72	0.021
Academic spend	1	0.002989	0.002989	0.34	0.565
Facilities spend	1	0.060855	0.060855	6.84	0.012
Error	43	0.382636	0.008899		
Total	46	0.533591			

```
Model Summary

        S    R-sq  R-sq(adj)  R-sq(pred)
0.0943319  28.29%     23.29%      15.28%


Coefficients

Term                  Coef   SE Coef  T-Value  P-Value   VIF
Constant             4.255     0.136    31.32    0.000
Ratio              -0.01351   0.00565    -2.39    0.021  1.25
Academic spend    -0.000022  0.000039    -0.58    0.565  1.18
Facilities spend   0.000161  0.000062     2.62    0.012  1.07


Regression Equation

Satisfaction = 4.255 - 0.01351 Ratio - 0.000022 Academic spend
+ 0.000161 Facilities spend


Fits and Diagnostics for Unusual Observations

Obs  Satisfaction     Fit    Resid  Std Resid
  4        4.2600  4.0750   0.1850       2.01  R
  7        3.9000  4.0949  -0.1949      -2.10  R
 39        4.0000  4.0284  -0.0284      -0.49    X

R  Large residual
X  Unusual X
```

You will notice that this output looks rather similar to that given for linear regression with one explanatory variable. Again, you can ignore a good deal of it. From the '`Regression Equation`' part of the output, you can see that the fitted multiple regression model is

$$\text{Satisfaction} = 4.255 - 0.01351\,\text{Ratio} - 0.000022\,\text{Academic spend} + 0.000161\,\text{Facilities spend}.$$

The Minitab output also includes the p-values for the individual two-sided tests of the null hypothesis $H_0 : \beta_j = 0$, for $j = 1, 2, 3$. These can be seen in the output in the table labelled '`Coefficients`' in the column labelled '`P-Value`'. Notice also that, as for linear regression with one explanatory variable, the end of Minitab's output still reports '`Fits and Diagnostics for Unusual Observations`', any individual data points that either have a large residual and/or are unusual in some other way: there are three of these values for this model.

(a) Explain why the data suggest that the regression coefficient for `Academic spend` is zero.

Because the data suggest that the regression coefficient for `Academic spend` is zero, it is sensible to try fitting another multiple regression model which just has `Ratio` and `Facilities spend` as explanatory variables. Do this now: fit another multiple regression model with `Satisfaction` as the response variable, `Ratio` and `Facilities spend` as explanatory variables, and obtain the residual and normal probability plots for this model.

(b) Write down the fitted multiple regression model.

(c) Explain why the data suggest that both of the regression coefficients are non-zero.

(d) Which universities have large residuals?

(e) Do the model assumptions seem reasonable?

4 Linearising relationships

This chapter is associated with Subsection 2.1 of Unit 12.

Sometimes it is possible to 'straighten out' or 'linearise' a non-linear relationship in a general regression model by a suitable transformation of the explanatory variable. Finding a suitable transformation can be tricky and it usually involves some trial and error. In this chapter, you will use Minitab to try various transformations in order to straighten out some non-linear relationships.

Activity 7 *Wind power*

An experiment was made into how the direct current output from a particular wind power generator changes with wind speed. A scatterplot of the response variable direct current output (current, for short) against the explanatory variable wind speed (in miles per hour) is shown in Figure 2.

Joglekar, G., Schuenemeyer, J.H. and LaRiccia, V. (1989) 'Lack-of-fit testing when replicates are not available', *American Statistician*, vol. 43, no. 3, pp. 135–43.

These data concern a small wind turbine, perhaps a predecessor of one like this

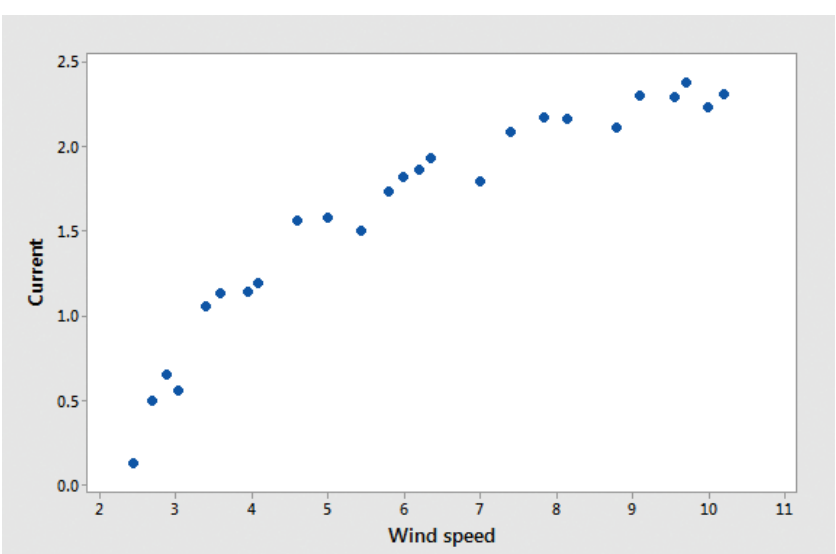


Figure 2 Direct current output against wind speed

The scatterplot suggests some sort of relationship between the current output and the wind speed, though not a linear relationship. The natural thing to do in a case such as this is to try to fit a curve to the data rather than a straight line. Analysis of these data by several statisticians has suggested that a reciprocal curve might describe the data, that is, the relationship between the variable 'current' and the transformed variable '1/wind speed' might be linear. So, letting Y_i denote the current and x_i the wind speed, one possible model for describing the relationship is

$$Y_i = \alpha + \beta\frac{1}{x_i} + W_i,$$

where the W_is are, as usual, independent normally distributed random variables with constant, zero mean and constant variance.

The worksheet **wind-power.mtw** contains the data from this experiment. Open the worksheet now. The variable `Wind speed` can be transformed and a scatterplot using the transformed data can be obtained, as follows.

- Choose **Calc** > **Calculator...** to open the **Calculator** dialogue box.
- To call the new variable 'Rspeed' ('R' is for 'Reciprocal'), type `Rspeed` in the **Store result in variable** field of the **Calculator** dialogue box.
- The wind speed readings are in the column `Wind speed` so enter `1/'Wind speed'` in the **Expression** field.
- Click on **OK** and the transformed data will be stored in a column named `Rspeed`.
- Now obtain a scatterplot of `Current` against the new variable `Rspeed` in the usual way, using **Graph** > **Scatterplot...**.

Such a scatterplot is shown in Figure 3.

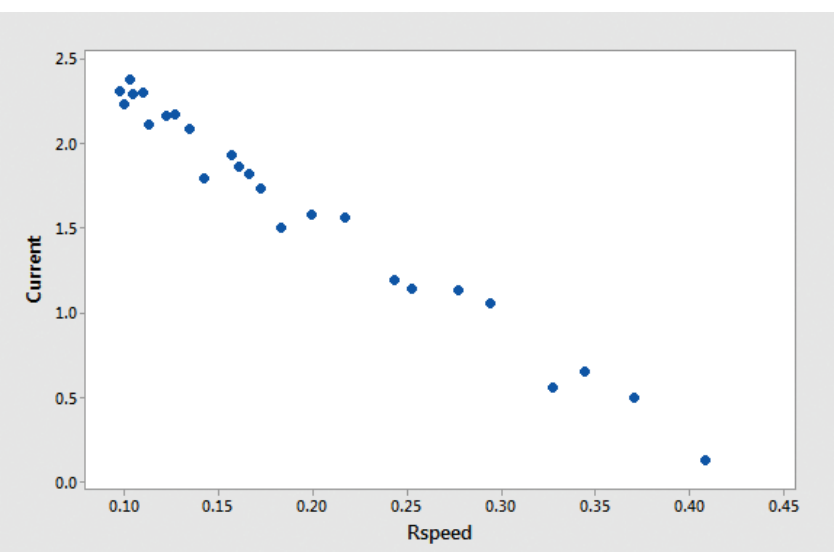


Figure 3 Direct current output against 1/wind speed

The scatterplot in Figure 3 suggests there might indeed be a straight-line relationship between the direct current output and the new explanatory variable 1/wind speed.

If it may be assumed that the random terms W_i are independent and normally distributed with constant, zero mean and constant variance, then a non-linear relationship between the response and explanatory variable has effectively been modelled by a linear regression model by transforming the explanatory variable.

Obtain the least squares line for the transformed data and check the assumptions of the fitted model.

Follow the procedure described in Chapter 2 of this computer book.

Activity 8 *Enzymatic reaction*

In this activity, you are asked to make several different transformations of a dataset in order to find an appropriate transformation to straighten out the non-linear relationship in the data from a chemistry experiment.

The data in the worksheet **enzymes.mtw** represent the velocity of an enzymatic reaction as a function of substrate concentration. Velocity is measured as counts per minute of radioactive product from the reaction; substrate concentration is measured in parts per million. Open the worksheet now.

Severini, T.A. (2000) *Likelihood methods in statistics*, Oxford, Oxford University Press.

A scatterplot of velocity against substrate concentration is shown in Figure 4.

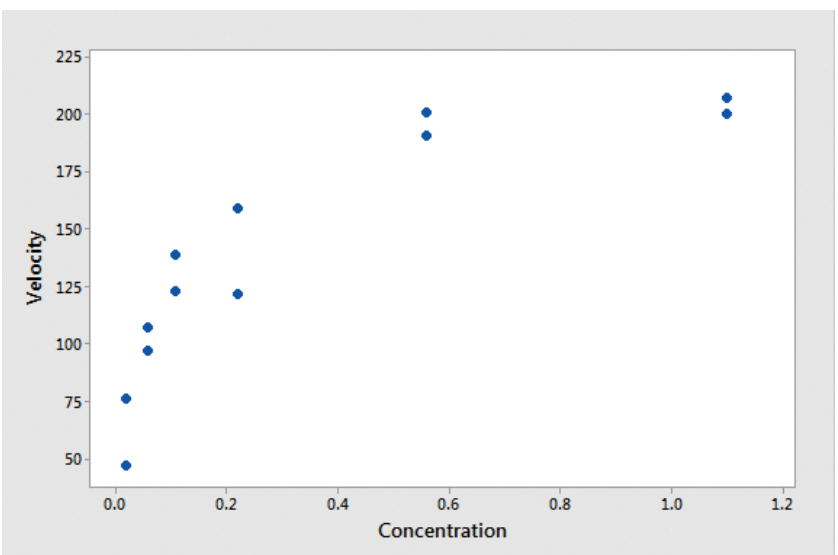


Figure 4 Velocity against substrate concentration

The data appear to follow a curve rather than a straight line: the velocity of the enzyme reaction increases steeply with concentration for low substrate concentrations, but more slowly for higher concentrations.

Three possible functions that might describe this curve are as follows (there may be others):

$$\sqrt{x} \qquad \log x \qquad 1/\sqrt{x}.$$

(a) For each of the functions in turn, use **Calc > Calculator...** to transform the data on substrate concentration to obtain values of a new explanatory variable, obtain a scatterplot of the velocity against the new explanatory variable, and comment on whether you think the transformation might be appropriate.

(b) Which of the three transformations would you choose to use?

(c) For the transformation that you chose, obtain the least squares regression line for the transformed data and check the assumptions of the fitted model. Do you stand by the preference that you stated in part (b)?

5 Modelling with Minitab

This chapter is associated with Section 4 of Unit 12.

In this chapter, you will undertake an extended exercise, or mini-project, in statistical modelling using Minitab. The chapter contains no new material, but aims to give you practice in the skills of statistical modelling. The project begins with some background and a scientific question, and the description of a dataset relevant to that question. Your task is to use these data to throw light on the original question. To do this, you will need to think about the question, the data, and what statistical models and methods to use.

The mini-project is quite long, so you may wish to do it in stages. In this case, you should save your work as you go as a Minitab project file. This will store all the intermediate variables you define, the output, and any graphs you create.

Example 1 *Michelson's determination of the speed of light*

The determination of the speed of light is a problem that has occupied the minds of philosophers, astronomers and physicists for centuries. Indeed, the very question as to whether light 'travels' at all, and therefore has a speed to measure, was not finally resolved until James Bradley's work on the parallax effect in the early eighteenth century. However, accurate methods for determining the speed of light did not become available until a century and a half later, pioneered by the physicist Albert Abraham Michelson (1852–1931). Michelson undertook a series of experiments that were to play an important role in the genesis of Albert Einstein's theory of relativity.

Michelson, A.A. (1880) 'Experimental determination of the velocity of light made at the U.S. Naval Academy, Annapolis', *U.S. Nautical Almanac Office Astronomical Papers*, vol. 1, no. 3, pp. 109–45.

Michelson undertook his first set of experiments in 1879. His experimental method was to shine a beam of light onto a rapidly rotating mirror, and to measure the displacement of the reflected beam due to the rotation of the mirror. The angular speed of the mirror was set using an electric tuning fork, itself calibrated against a standard tuning fork after correcting for the ambient temperature. Michelson measured the displacement of the reflected light beam using a precision micrometer. He then converted displacement to speed in units of kilometres per second ($\mathrm{km\,s^{-1}}$). He undertook 100 sets of experiments over several days (morning and afternoon) between 5 June and 2 July 1879. In each set of experiments he made ten speed determinations and reported the average of the ten values.

Albert Michelson and Albert Einstein met in 1931, shortly before Michelson's death. They are pictured at the Mount Wilson Observatory, California, USA. From left to right are Milton Humason, Edwin Hubble, Charles St. John, Michelson, Einstein, W.W. Campbell and Walter S. Adams

Today, very much more accurate methods for determining the speed of light are used than were available to Michelson. The speed of light in a vacuum is now known to be 299 792 $\mathrm{km\,s^{-1}}$ (to the nearest $\mathrm{km\,s^{-1}}$) while the speed of light in air – the value being measured by Michelson – is around 299 703 $\mathrm{km\,s^{-1}}$. The questions of interest are as follows.

- How close to the true value of the speed of light in air was Michelson's 1879 estimate?
- What factors might have affected his measurements?

In this chapter, you will consider the two questions identified in Example 1 and their setting from a statistical modelling point of view, and explore statistical models to answer them.

The background description contains several technical terms which may be unfamiliar to you. This is a common situation for a statistician: you will often be working with specialists in the relevant field of application, to whom you can turn for advice and further information. In this case, the salient features are that Michelson's determinations of the speed of light were made at different times of day on different days, using a rather complicated apparatus, and that various factors such as temperature were allowed for. For the purposes of this exercise, you should not worry if you do not know what the parallax effect is, or how a precision micrometer works!

Activity 9 *Thinking about the setting*

Re-read the above background in the light of the questions of interest. From the brief description given, what factors do you think might have had an effect on the accuracy of Michelson's measurements? Why did Michelson repeat his experiment so many times?

Activity 10 *Thinking about the question*

The data that Michelson reported are measurements (or more precisely, averages of ten measurements) of the speed of light in air in $\mathrm{km\,s^{-1}}$ made using his apparatus. He also reported various other features of his

experiment, such as the day and time of day the measurements were made, and the ambient temperature.

(a) Consider the question 'How close to the true value of the speed of light in air was Michelson's 1879 estimate?' How might you answer that question?

(b) It is perhaps unlikely that Michelson's estimate of the speed of light in air would exactly match the true value, because of measurement error. However, it would be interesting to know whether Michelson's estimate is consistent with the true value, allowing for the variability due to measurement error. What statistical techniques might you use to address this?

In part (b) of Activity 10 the question in part (a) was refined, from 'How close to the true value is Michelson's estimate?' to 'Is Michelson's estimate consistent with the true value?' Refining the question is an important part of a statistical analysis: the aim is always to extract as much information as possible of relevance to the problem, rather than answer the question in a purely literal sense.

In Activity 10, you set out a general approach to the problem. Before you look at the data, it is worth thinking in a little more detail about how you might implement this approach. In particular, you might reflect on what population Michelson's data are drawn from by imagining a hypothetical population of all possible measurements using the method described. Michelson's data then represent a sample of size $n = 100$ from this population.

Activity 11 *Thinking about the data*

Would you model Michelson's measurements using a discrete or a continuous variable? From what you know about the setting of the experiment, identify a possible model for these data.

Now we are ready to take a look at the data.

Activity 12 *Exploring the data*

These data are taken from MacKay, R.J. and Oldford, R.W. (2000) 'Scientific method, statistical method and the speed of light', *Statistical Science*, vol. 15, no. 3, pp. 254–78.

Michelson's data are contained in the Minitab worksheet **michelson.mtw**. Open the worksheet now. The worksheet contains the following four variables.

`Speed`	speed of light in air, in kilometres per second ($\mathrm{km\,s^{-1}}$)
`Day`	day of the experiment (5 June = 1)
`Time`	time of day (AM = 1 hour after sunrise, PM = 1 hour before sunset)
`Temperature`	air temperature, in degrees Fahrenheit.

(a) Obtain a histogram of the variable `Speed`. Is the model you identified in Activity 11 likely to be a reasonable one?

(b) Use a suitable graphical method to check your model for the variable `Speed`. If necessary, revise your model.

Activity 13 *Calculations*

Having established a suitable model in Activities 11 and 12, the next step in the analysis is to use this model to carry out the procedure you identified in Activity 10(b). This involves calculating the mean speed, calculating a confidence interval for the mean, and carrying out a hypothesis test.

(a) Calculate the sample mean of the speed of light in air measurements in $\mathrm{km\,s^{-1}}$, and obtain 95% and 99% confidence intervals for the population mean.

(b) The true value of the speed of light in air is around $299\,703\ \mathrm{km\,s^{-1}}$. Comment on how close Michelson's result was to the true value.

(c) Carry out a hypothesis test of the hypotheses

$$H_0 : \mu = 299\,703, \quad H_1 : \mu \neq 299\,703,$$

where μ is the population mean of Michelson's measurements.

(d) What do you conclude about the discrepancy between the mean of Michelson's measurements and the true value of the speed of light in air? Is the discrepancy likely to be due to random measurement error? What other explanations might there be?

In Activity 13 you saw that the true value of the speed of light in air lies outside the 99% confidence interval for the mean calculated from Michelson's data. A hypothesis test provided strong evidence against the null hypothesis that the underlying mean of Michelson's measurements is equal to the speed of light in air. Michelson's values seem to be systematically higher than the true value. Since the speed of light in air is unlikely to have changed over time (at least, since the late nineteenth century!), we conclude that there was probably some systematic bias in Michelson's experiment.

This is in metres per second rather than kilometres per second

The second question for investigation is therefore especially pertinent: 'What factors might have affected Michelson's measurements?' To answer this, a detailed examination of Michelson's apparatus would be required. However, we have data on day, time of day, and ambient temperature. Thus we can at least make a start on the problem by looking at the refined question 'What effect did day, time of day, and ambient temperature have on Michelson's experiment?'

Activity 14 *The effect of day of experiment*

(a) Using a suitable graphical display, investigate whether the speed measurements varied with the day of the experiment.

(b) In order to investigate further, fit a regression line to the data used in part (a). What is the equation of the least squares fitted line? Produce appropriate plots to check the assumptions of the linear regression model for these data.

(c) The p-value associated with the test of the null hypothesis $H_0 : \beta = 0$, that is, of no relationship between speed measurement and day, can be found in the Minitab output from the analysis of part (b) in the Session window of Minitab: look at the table headed `Coefficients` and take the `P-Value` in the row labelled `Day`. Using this p-value, what conclusion do you draw about whether or not there is a dependence of Michelson's speed measurements on the day on which the experiment was performed?

(d) According to the fitted linear regression model, what is the average effect on Michelson's measurements of each passing day?

Activity 15 *The effect of temperature*

Michelson adjusted his results for the effect of temperature. In this activity you are asked to examine whether this adjustment was effective, that is, whether it removed all the effects of temperature variation on the experimental results.

(a) Use a suitable graph to investigate visually the relationship between speed of light measurements and ambient temperature. What is your impression?

(b) Fit a regression line to the data used in part (a). What is the equation of the least squares fitted line? Produce appropriate plots to check the assumptions of the linear regression model for these data.

(c) Obtain the p-value associated with the test of the null hypothesis $H_0 : \beta = 0$, that is, of no relationship between speed measurement and temperature. What conclusion do you draw about whether or not, despite his efforts to adjust for temperature, there remains a dependence of Michelson's speed measurements on the temperature at which the experiment was performed?

(d) According to the fitted linear regression model, what is the average effect on Michelson's measurements of a 1 °F increase in temperature?

Activity 16 *The effect of time of day*

Michelson also recorded the time of day of his experiments, a feature that is also likely to be related to temperature.

Missing values in a text variable record the word 'Missing'. In contrast, missing values in numeric variables are coded $*$.

(a) The variable `Time` is a text variable, taking the values `AM` for morning and `PM` for afternoon. Some of these values are missing. Find out how many missing values there are. Hint: look in the Project Manager window.

(b) Obtain two comparative boxplots, one to display the relationship between speed of light measurement and time of day and the other to display the relationship between temperature and time of day. To obtain these comparative boxplots, after choosing **Graph** > **Boxplot**, select **One Y, With Groups** in the **Boxplots** dialogue box. For the first boxplot, enter `Speed` in the **Graph variables** field; for the second boxplot, enter `Temperature` in this field. For each boxplot, `Time` is entered in the **Categorical variables for grouping (1–4, outermost first)** field. You can stop Minitab displaying a boxplot corresponding to the two missing values of `Time` as follows.

Also, ensure that **Transpose value and category scales** is ticked in the **Boxplot: Scale** dialogue box in order to obtain horizontal boxplots.

- Click on **Data Options...** to open the **Boxplot: Data Options** dialogue box.
- Click on the **Group Options** tab to view the **Group Options** panel.
- Deselect **Include missing as a group**.
- Click on **OK** to close this dialogue box, then click on **OK** to produce the comparative boxplot.

What do you observe?

(c) Use the following instructions to obtain a normal probability plot of the speed of light measurements that were made in the morning.

- Obtain the **Probability Plot: Single** dialogue box in the usual way from **Graph** > **Probability Plot...**, with `Speed` entered in the **Graph variables** field.
- Click on **Data Options...** to open the **Probability Plot: Data Options** dialogue box.
- Ensure that the **Subset** panel is uppermost.
- Under **Specify Which Rows To Include**, select **Rows that match**.
- Click on the **Condition...** button.
- In the **Probability Plot: Subset** dialogue box, enter `Time = "AM"` in the **Condition** field. Note: it is important to use double quotes around `AM` otherwise this step will fail.
- Click on **OK** to close this dialogue box, then on **OK** again to close the next dialogue box, and finally click on a third **OK** to produce the required normal probability plot.

In a similar way, obtain a second normal probability plot of the speed of light measurements that were made in the afternoon. Are you satisfied with the normality of the speed of light measurements in each group?

(d) Assuming normality of the speed of light measurements in both the morning and the afternoon, check the other assumption (of equality of variances; see Subsection 4.4 of Unit 8) that will allow you – if satisfied – to proceed to obtain a 95% confidence interval for the mean difference between the speed of light measurements in the morning and the speed of light measurements in the afternoon.

(e) Obtain the required confidence interval. To do so, use **Stat** > **Basic Statistics** > **2-Sample t...**. In the **Two-Sample t for the Mean** dialogue box, make sure that **Both samples are in one column** is selected from the drop-down list at the top, and then put `Speed` in the **Samples** field and `Time` in the **Sample IDs** field. Also, in the **Two-Sample t: Options** dialogue box, make sure that the **Confidence level** is set to 95 and select **Assume equal variances**. What does the confidence interval tell you?

(f) In one sentence, relate the outcome of this activity to that of the previous activity.

Exercises

Exercise 1 *Coaching in, and membership of, sports clubs*

In Activity 25 of Unit 1, data were considered on the percentages of adults who were members of sports clubs in 49 areas in England in 2014–2015 and the percentages of adults in those areas who received coaching in a sport in the previous year. This exercise concerns the dependence of the percentages of adults receiving coaching on the percentage of adults who were members of sports clubs. Figure 5 is a scatterplot of these values that is repeated from Figure 29 of Unit 1. It was concluded in the solution to Activity 25 of Unit 1 that a fairly strong, positive, linear relationship existed between the variables, albeit with some potential outliers.

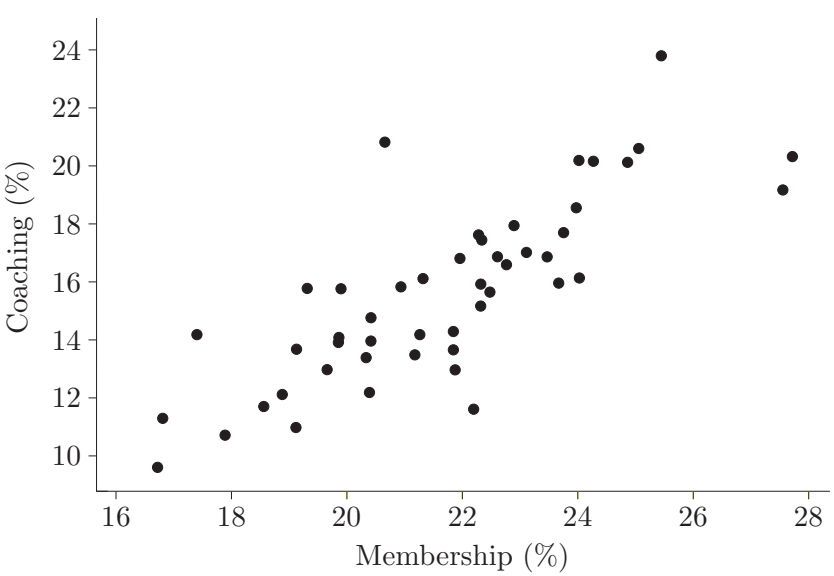


Figure 5 Sports coaching against sports club membership

The data are given in the worksheet **coaching.mtw**. Open the worksheet now. The response variable is in the column `Coaching` and the explanatory variable is in the column `Membership`.

(a) Calculate the least squares line for the relationship between coaching and membership.

Stat > Regression > Regression > Fit regression model...

(b) Interpret what the values of the least squares estimates of the parameters of the regression line tell us.

(c) Check the assumptions of the fitted model. Are the assumptions reasonable for these data?

Exercise 2 *Antique grandfather clocks*

The Minitab worksheet **clocks.mtw** contains data from the early 1990s on the selling price (in pounds) of $n = 32$ antique longcase/grandfather clocks along with the age of the clock (in years) and the number of bidders participating in the sale at which the clock was sold. Interest lay in how the selling price, the response variable Y, depends on the two explanatory variables, the age of the clock (x_1) and the number of bidders (x_2).

Mendenhall, W. and Sincich, T.L. (1993) *A Second Course in Statistics: Regression Analysis*, 6th edn, Prentice-Hall, p. 173.

Open this worksheet now. You will find the selling prices in the column labelled `Price`, the ages in the column labelled `Age` and the numbers of bidders in the column labelled `Bidders`. (It is unclear why the selling prices are not in more rounded numbers of pounds, e.g. £950, £1100 etc., as you might expect.)

Fit a multiple linear regression model using both explanatory variables; as you do so, also obtain a residual plot and a normal probability plot of the residuals.

(a) What is the equation of the fitted multiple regression model?

(b) Explain why the data suggest that both of the regression coefficients are non-zero.

(c) Interpret the regression coefficients.

(d) In the early 1990s, I had a 150-year-old grandfather clock that I wished to sell. The auctioneer expected 6 bidders for the clock at the sale. Assuming this was the case, according to the fitted model, what price might I have hoped to get for my clock?

(e) Do the model assumptions seem reasonable?

Exercise 3 *The evolution of human brain size*

The data in the worksheet **brainsize.mtw** represent the cranial capacity (brain size, in cubic centimetres) of human fossil skulls and the ages of the skulls (in Ka, that is, thousands of years). Open the worksheet now.

Lee, S.-H. and Wolpoff, M.H. (2003) 'The pattern of evolution in Pleistocene human brain size', *Paleobiology*, vol. 29, no. 2, pp. 186–96.

(a) Obtain a scatterplot of cranial capacity against age. Comment on what you see.

Modern (top) and ancient (bottom) human skulls (Figure 1 of Lee and Wolpoff, 2003)

(b) Three possible functions that might describe the relationship between cranial capacity and age are as follows (there may be others):

$$\sqrt{x} \qquad \log x \qquad 1/x.$$

(i) Using each of the functions in turn, transform the data on age to obtain values of a new explanatory variable, obtain a scatterplot of the cranial capacity against the new explanatory variable and comment on whether you think the transformation might be appropriate.

(ii) Which of the three transformations would you choose?

(iii) For the transformation that you chose, obtain the least squares regression line for the transformed data and check the assumptions of the fitted model. Do you stand by the preference that you stated in part (b)(ii)?

Solutions to activities

Solution to Activity 2

(a) When the points on a scatterplot lie roughly along a line from the lower-left corner to the upper-right corner of the plot, adding a point in the upper-left (or lower-right) corner of the diagram causes the least squares line to become less steep: the line tilts towards the extra point. This is illustrated in Figure 6, which shows two scatterplots produced by the animation. In Figure 6(a), the initial points are shown together with the least squares line; in Figure 6(b), the same points are shown with an additional point at the upper-left corner, together with the new least squares line.

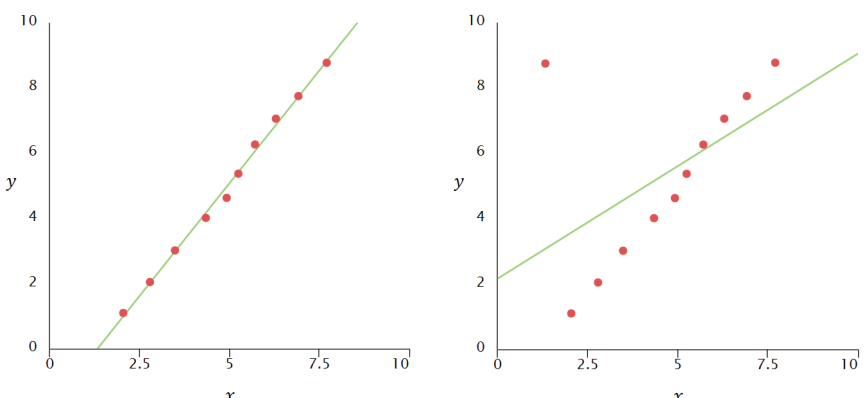


Figure 6 Points on a scatterplot together with the least squares lines (a) without, and (b) with, an additional upper-left point

When an outlying point is added at the top (or at the bottom) of the scatterplot halfway across the diagram, the least squares line shifts slightly up or down towards the new point, but the slope of the line does not change much. Figure 7(a) (overleaf) is a repeat of Figure 6(a), while Figure 7(b) shows the same points with an additional point at the top halfway across the diagram, together with the new least squares line.

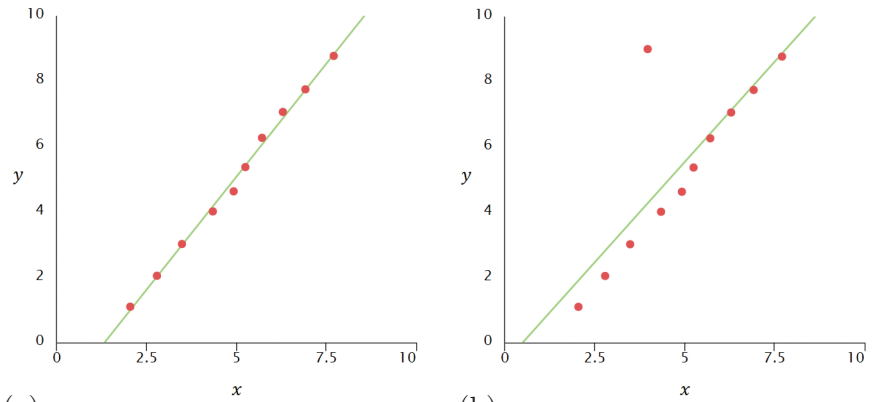


Figure 7 Points on a scatterplot together with the least squares lines (a) without, and (b) with, an additional point at the top halfway across

(b) When a scatterplot contains a cluster of points on the left-hand side of the plot and a single outlying point on the right-hand side, even a small change up or down in the position of the outlier changes the least squares line drastically. This is illustrated in Figure 8.

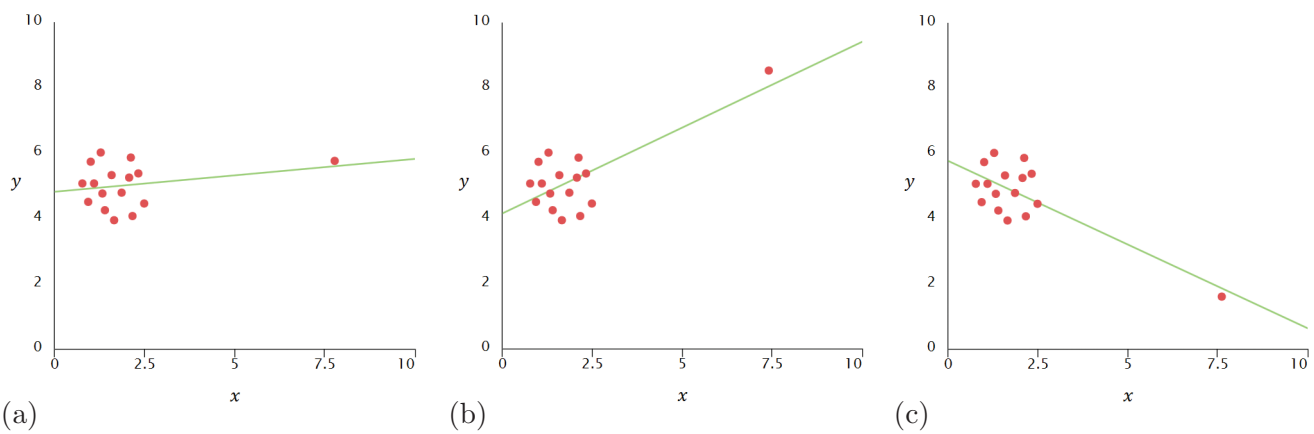


Figure 8 Cluster of points with an additional point in (a) its original position, (b) shifted up, and (c) shifted down, together with the least squares lines

Solution to Activity 4

(a) Using **Graph** > **Scatterplot...**, a scatterplot of race time against wind speed can be obtained and is shown in Figure 9.

The scatterplot shows a general downward pattern. There is a lot of scatter, but it is not unreasonable to describe the pattern as roughly linear. A linear regression model might be appropriate for these data.

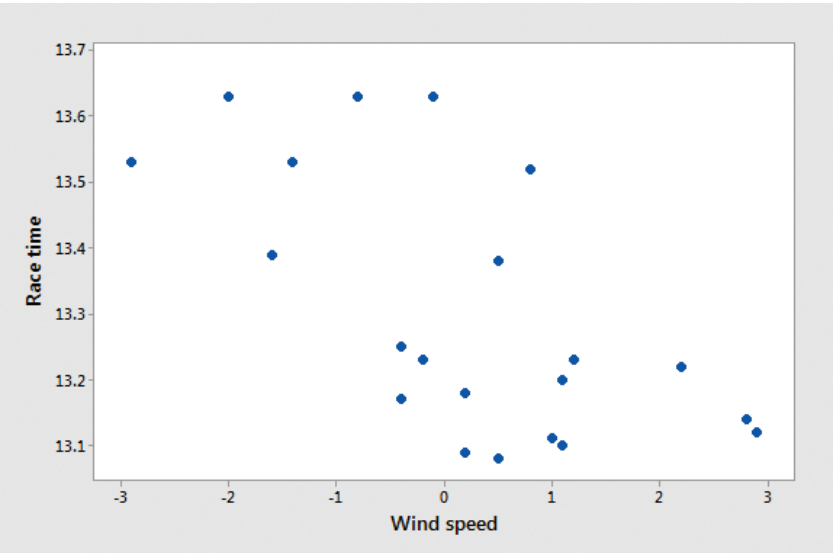


Figure 9 Race time against wind speed

(b) Choose **Stat** > **Regression** > **Regression** > **Fit Regression Model...** with `Race time` in the **Responses** field and `Wind speed` in the **Continuous predictors** field. The least squares line for the data is given by Minitab as

$$\text{Race time} = 13.3218 - 0.0846\,\text{Wind speed}.$$

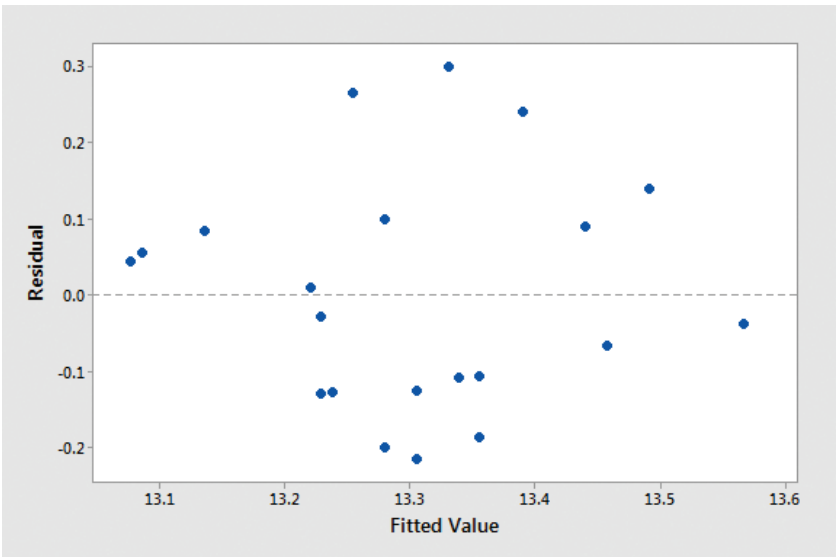
(c) The value $\widehat{\alpha} = 13.3218$ is the estimated value of the intercept: the value of the regression line when $x = 0$. It is very meaningful in this case as the estimated value of the race time (in seconds) that, on average, Colin Jackson would have achieved in 1990 if there were no wind at all.

The value $\widehat{\beta} = -0.0846$ is the estimated value of the slope. It estimates that, for each additional metre per second of following wind, Colin Jackson, in 1990, might on average have been able to decrease his race time by 0.0846 seconds. (Notice that $\widehat{\beta}$ has a negative sign, hence the decrease in the response variable for an increase in the value of the explanatory variable.)

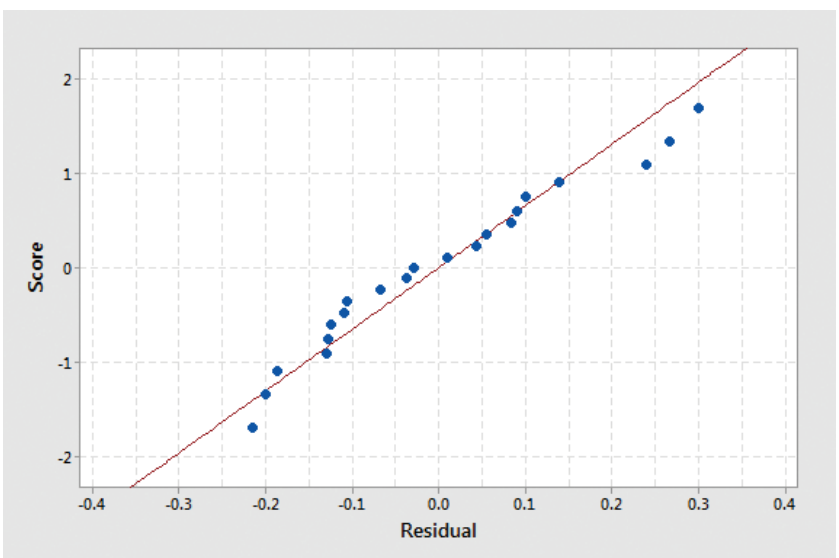
(d) No unusual points are highlighted in the Minitab output in the Session window.

The residual plot and normal probability plot of the residuals are obtained by selecting **Normal probability plot of residuals** and **Residuals versus fits** in the **Regression: Graphs** dialogue box when calculating the least squares line (as in part (b)). The resulting plots are shown in Figure 10 (overleaf).

If your plots are slightly different to these, check that **Residuals for Plots** is set to **Regular** in the **Regression: Graphs** dialogue box.



(a)



(b)

Figure 10 (a) Residual plot (b) normal probability plot of residuals

There is no very obvious pattern in the residual plot in Figure 10(a), so the assumption of constant, zero mean and constant variance of the random terms seems appropriate. That said, it might be argued that the variance is not constant, being smaller for both small and large fitted values than for the many more moderate ones. However, the effect is not strong and could, at least partially, be explained by the smaller numbers of observations with extreme fitted values compared with those with moderate fitted values. This is because small numbers of observations give less opportunity for the consequences of large variability to be visible.

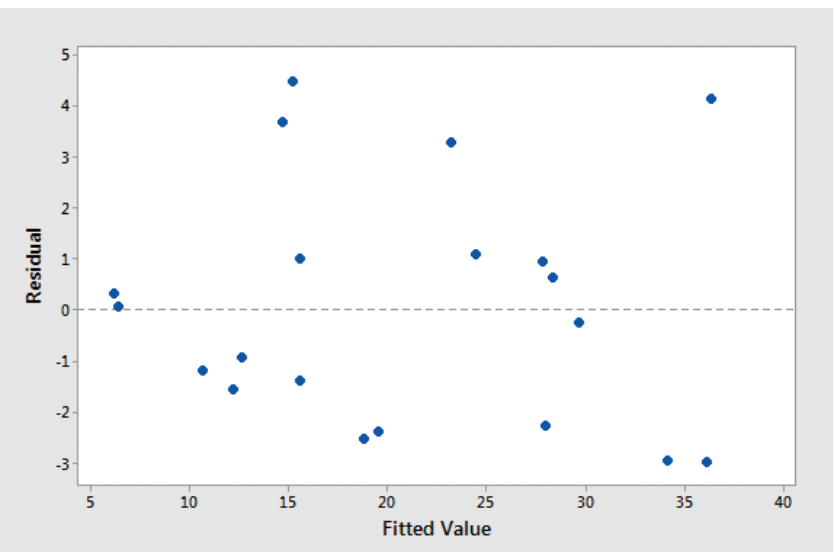
The points in the normal probability plot in Figure 10(b) lie roughly along a straight line, so the assumption of normality of the random terms seems plausible.

All told, many statisticians would be comfortable with the assumptions underlying the linear regression model for these data.

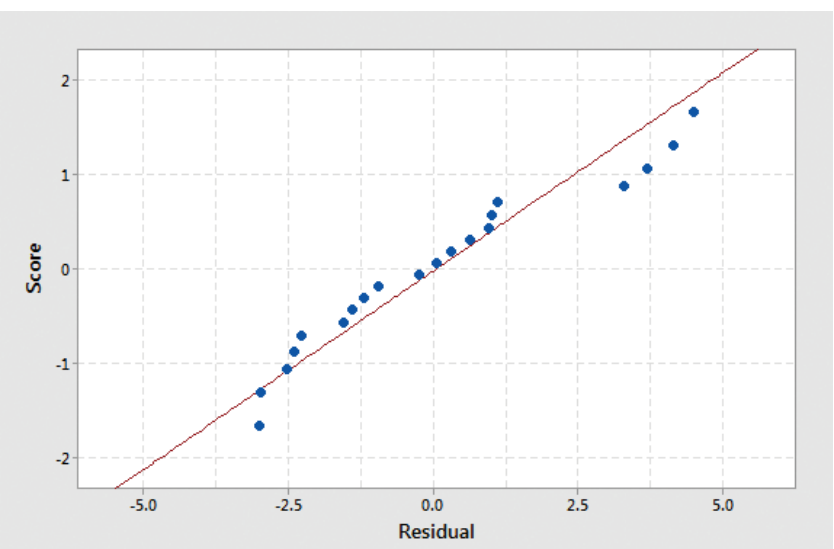
Solution to Activity 5

No unusual points are highlighted in the Minitab output in the Session window.

The residual plot and the normal probability plot of the residuals produced by Minitab are shown in Figure 11.



(a)



(b)

Figure 11 (a) Residual plot (b) normal probability plot of residuals

The residuals in Figure 11(a) seem to fluctuate around zero in an unsystematic fashion, so the assumption of constant, zero mean and constant variance of the random terms seems reasonable. The points in the normal probability plot in Figure 11(b) form something of a wavy line, though they probably lie close enough to a straight line to suggest that a normal model for the random terms of the road distance data is plausible. Therefore, a linear regression model with the constraint that the straight line through the data goes through the origin might be a good model for these data.

Solution to Activity 6

(a) The p-value associated with the regression coefficient for `Academic spend` is 0.565. This is much greater than 0.1 and so there is little or no evidence to suggest that this regression coefficient is non-zero.

(b) The fitted multiple regression model is

$$\text{Satisfaction} = 4.2006 - 0.01222\,\text{Ratio} + 0.000166\,\text{Facilities spend}.$$

(c) The output table labelled '`Coefficients`' is reproduced below.

```
Coefficients
Term                  Coef   SE Coef  T-Value  P-Value   VIF
Constant            4.2006    0.0973    43.19    0.000
Ratio              -0.01222   0.00515    -2.37    0.022  1.06
Facilities spend  0.000166  0.000061     2.72    0.009  1.06
```

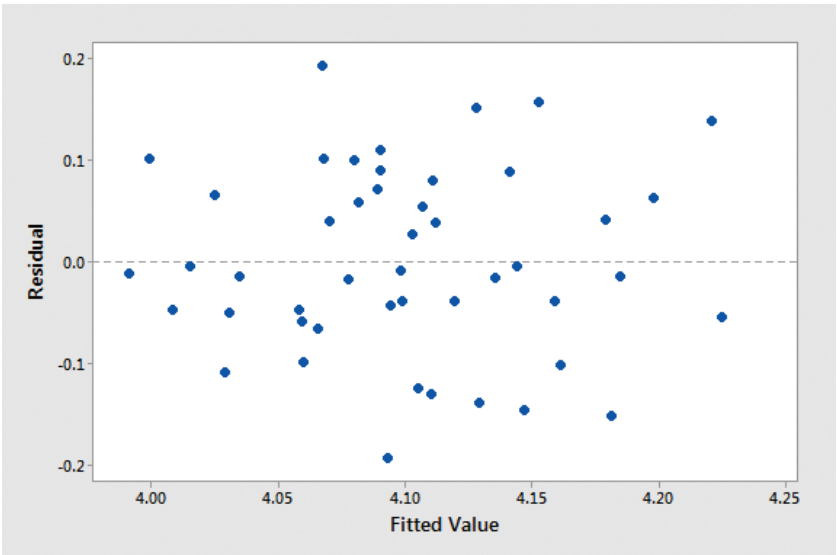
From this, the p-value associated with the regression coefficient for `Ratio` is 0.022, and since $0.01 < 0.022 < 0.05$, there is moderate evidence to suggest that this regression coefficient is non-zero. The p-value associated with the regression coefficient for `Facilities spend` is 0.009, and since $0.009 < 0.01$, there is strong evidence to suggest that this regression coefficient is non-zero.

(d) The observations which are unusual or have large residuals are identified at the end of Minitab's output as reproduced below.

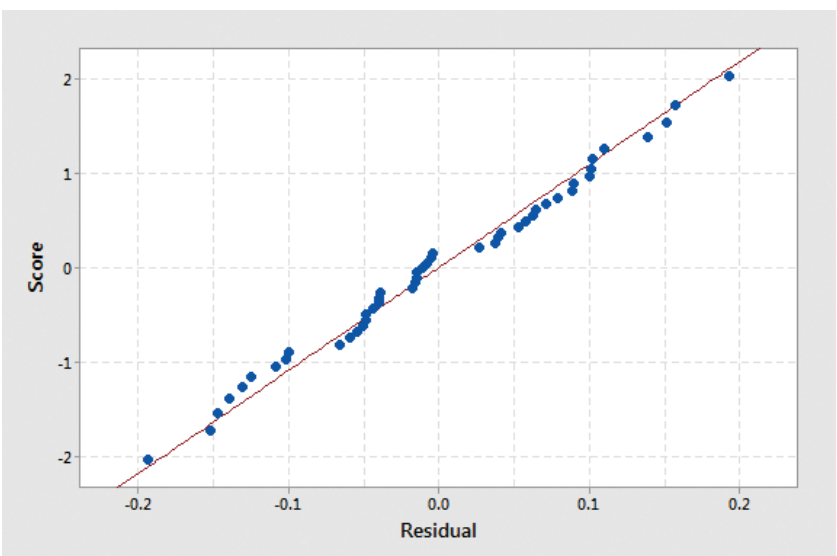
```
Fits and Diagnostics for Unusual Observations

Obs  Satisfaction     Fit    Resid  Std Resid
  4        4.2600  4.0672   0.1928       2.09  R
  7        3.9000  4.0932  -0.1932      -2.09  R

R  Large residual
```

From this output, there are two large residuals: a positive one associated with observation 4, and a negative one, associated with observation 7. These universities are Bangor and Bournemouth, respectively.

(e) The residual plot and normal probability plot of residuals are shown in Figure 12.



(a)



(b)

Figure 12 (a) Residual plot (b) normal probability plot of residuals

The points in the residual plot seem to be fairly randomly scattered about zero, and so the assumption that the random terms have constant, zero mean and constant variance seems reasonable. The two 'large residuals' notified by Minitab hardly stand out on this plot at all.

Most of the points in the normal probability plot generally follow a straight line and so the assumption that the random terms are normally distributed is reasonable.

Solution to Activity 7

Use **Stat** > **Regression** > **Regression** > **Fit Regression Model...** with `Current` in the **Responses** field and `Rspeed` in the **Continuous predictors** field, obtaining the required graphs – a residual plot and a normal probability plot of residuals, for checking the assumptions – by clicking on **Graphs...** to obtain the **Regression: Graphs** dialogue box and selecting **Normal probability plot of residuals** and **Residuals versus fits**.

The least squares line for the transformed data is

$$\text{Current} = 2.9789 - 6.935\,\text{Rspeed},$$

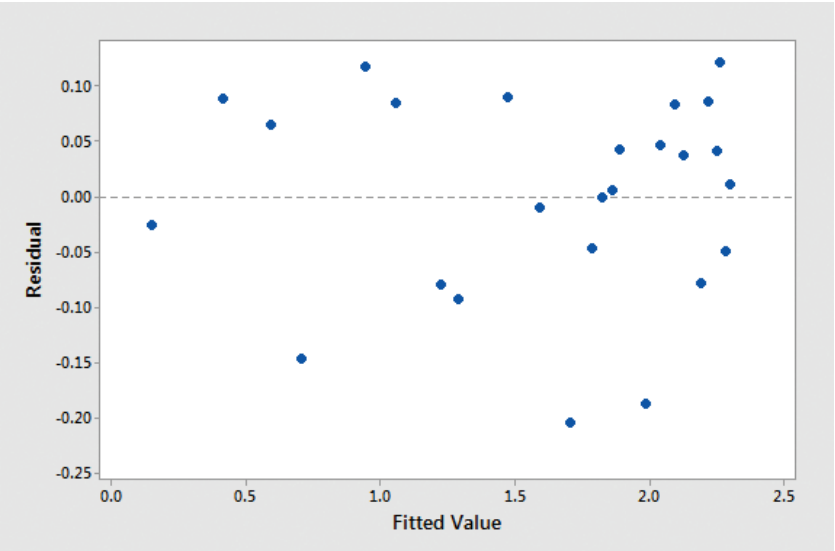
or

$$\text{Current} = 2.9789 - \frac{6.935}{\text{Wind speed}}.$$

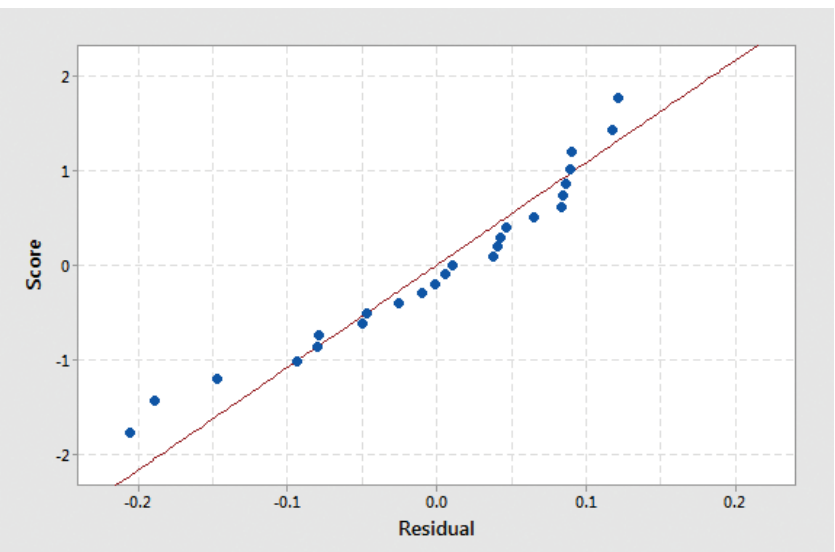
A residual plot and a normal probability plot of the residuals are shown in Figure 13.

There is no obvious pattern in the residual plot, nor does the spread of the residuals vary with the fitted values in a systematic way, so the assumptions of constant, zero mean and constant variance of the random terms seems appropriate.

There is, however, some curvature in the normal probability plot suggesting that the assumption of normality of the random terms might not be valid in this case. The two 'large residuals' identified by Minitab in the Session window are the two points to the bottom-left of the normal probability plot. These are also rather out of line with the normality assumption.



(a)



(b)

Figure 13 (a) Residual plot (b) normal probability plot of the residuals

Solution to Activity 8

(a) Square roots and natural logarithms are included in the list of functions available using **Calc** > **Calculator...**; select `Square root` and `Natural log (log base e)`, respectively. The three scatterplots will be as shown in Figures 14, 15 and 16 (overleaf).

In Figure 14, velocity is plotted against the square root of concentration. The points seem to follow a curve rather than a straight line. So a square root transformation of the data on substrate concentration is not appropriate.

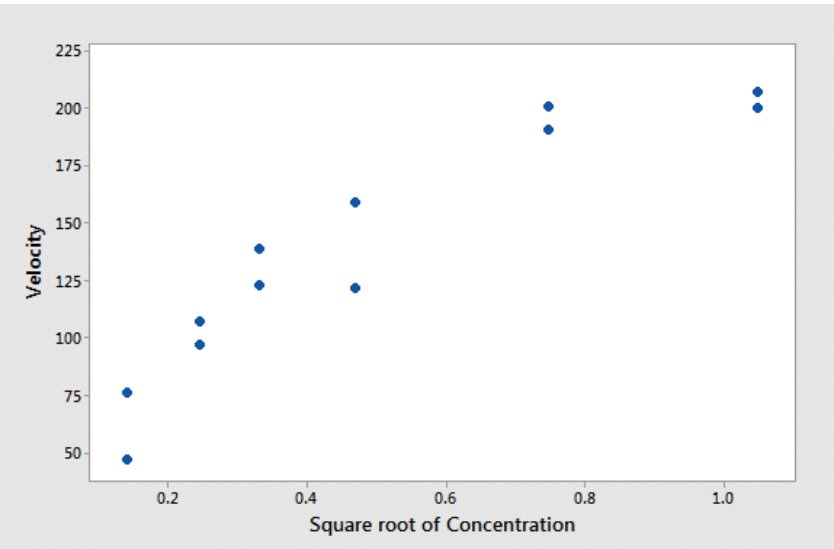


Figure 14 Velocity against $\sqrt{\text{concentration}}$

In Figure 15, velocity is plotted against the log of concentration. The points lie fairly close to a straight line. It seems that a logarithmic transformation might be appropriate for straightening out the non-linear relationship in the original data.

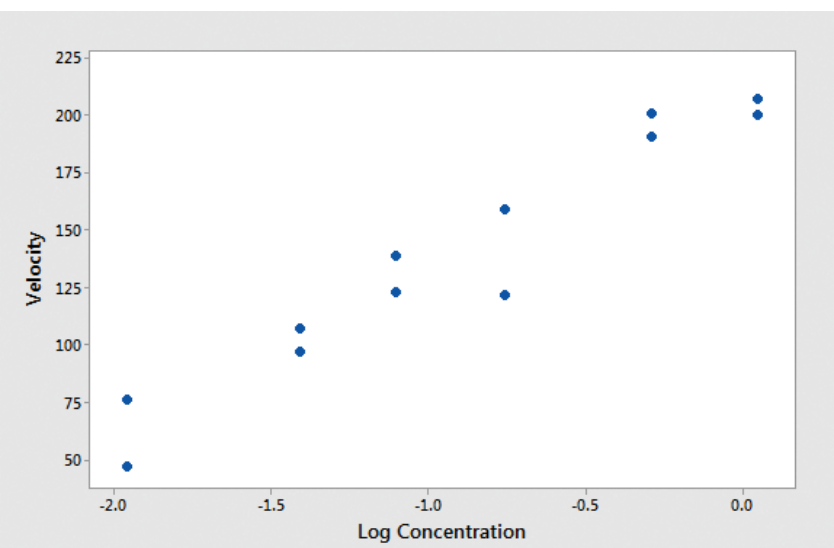


Figure 15 Velocity against log(concentration)

In Figure 16, velocity is plotted against the reciprocal of the square root of concentration. The points lie roughly along a straight line, although there can be argued to be some slight curvature for large values of the explanatory variable (a straight line that fits the main body of points to the left will 'undershoot' the two points to the right). The transformation $1/\sqrt{x}$ might be appropriate for straightening out the non-linear relationship in the original data.

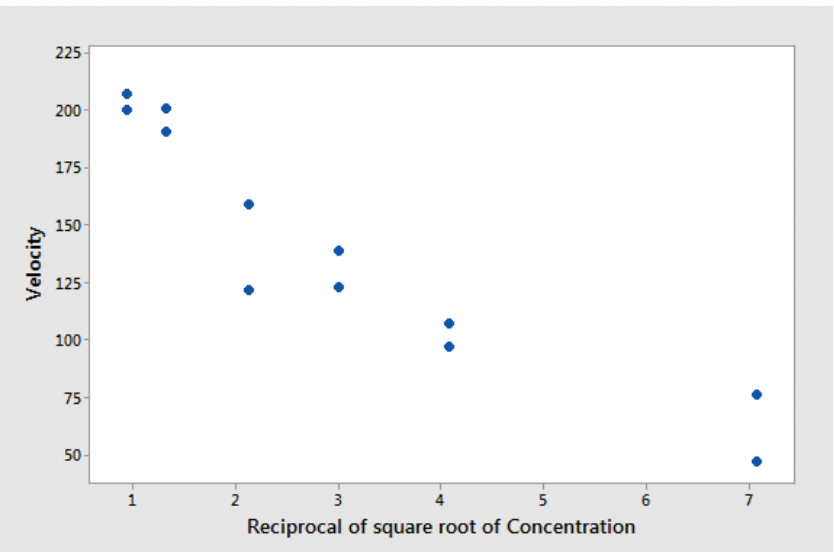


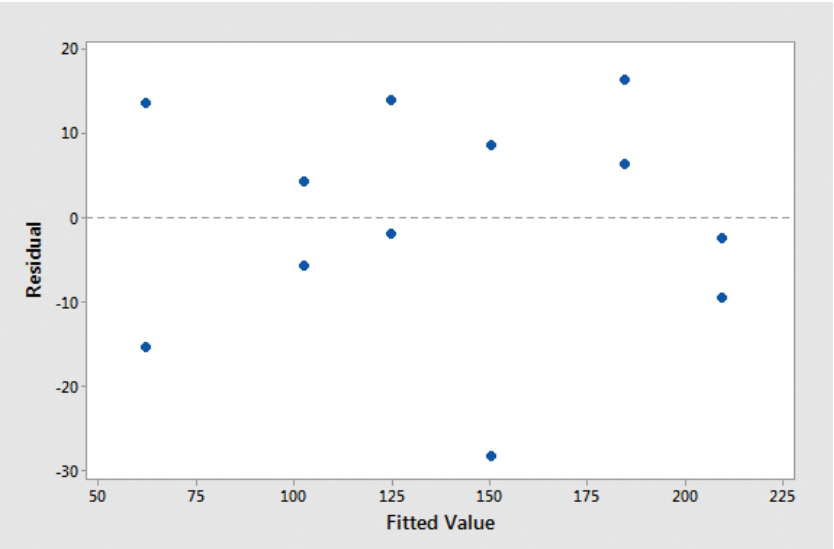
Figure 16 Velocity against $1/\sqrt{\text{concentration}}$

(b) Both the second and third transformations seem to straighten out the non-linear relationship reasonably well, so either could reasonably be chosen, whereas the square root transformation does not seem to be appropriate.

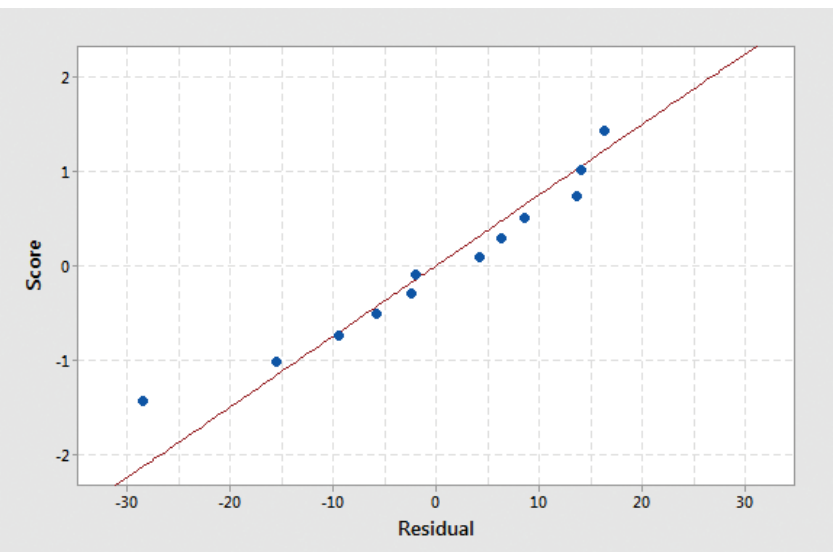
(c) If you opted for the second transformation, log, then the equation of the least squares line for the transformed data would be

$$\text{Velocity} = 205.92 + 36.68 \log(\text{Concentration}).$$

A residual plot and normal probability plot of the residuals are shown in Figure 17 (overleaf).



(a)



(b)

Figure 17 (a) Residual plot (b) normal probability plot of the residuals; log transformation

The points in the residual plot appear to be scattered randomly on the plot so the assumption of constant, zero mean and constant variance of the random terms seems appropriate. The points in the normal probability plot lie fairly close to a straight line so the assumption of normality of the random terms is plausible.

If you opted for the third transformation, the reciprocal of the square root, then the equation of the least squares line for the transformed data would be

$$\text{Velocity} = 210.3 - \frac{22.99}{\sqrt{\text{Concentration}}}.$$

A residual plot is shown in Figure 18.

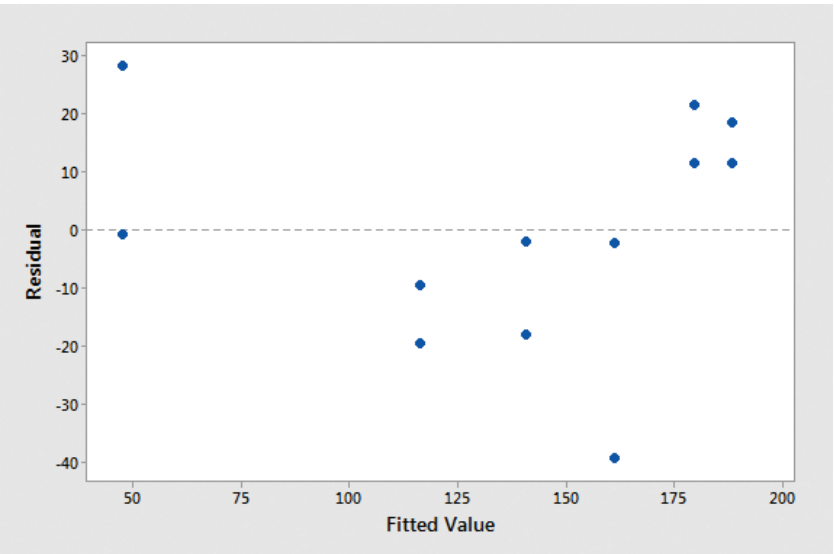


Figure 18 Residual plot; reciprocal of square root transformation

Although the sample size is small, there is a clear suggestion of a curved pattern in the residual plot, suggesting that the assumption of constant, zero mean of the random terms, and hence linearity of the mean of the response variables, is not particularly reasonable for this transformation.

A normal probability plot is not considered because normality is not of interest if the random terms have non-constant mean.

There is reasonable support for the logarithmic transformation in this case. However, you might argue that there is insufficiently clear structure in the residual plot and too few data points to be dogmatic about this.

Solution to Activity 9

Many different factors might have had an effect on the accuracy of the measurements. Measurement errors might have affected the determination of the micrometer displacements, and might have varied according to the day and time of day of the experiment. Variation in the settings used for the angular speed of the mirror and other features of the apparatus might have introduced some variability. Michelson's equipment responded differently according to the ambient temperature, a factor that was taken into account in the calculations. However, the effect of temperature might not have been completely accounted for.

Michelson undertook 100 sets of ten experiments (producing 100 averages of ten speed determinations) in order to reduce the effect of measurement error. Repeating the experiment many times will reduce the variance of the sample mean; however, it will not correct systematic biases.

Solution to Activity 10

(a) Michelson's estimate of the speed of light in air is obtained by calculating the mean of his 100 measurements. The question can then be answered by comparing this estimate with the true value of the speed of light, which is accurately known today.

(b) As stated in the question, it is rather unlikely that Michelson's estimate of the speed of light in air would exactly match the true value. A more relevant question is whether the confidence interval for the mean, obtained from Michelson's experiments, contains the true value. Alternatively, you could test the hypothesis that Michelson's estimate comes from a population with mean equal to the true speed of light in air. So, confidence intervals and hypothesis tests seem to be appropriate statistical techniques to use.

Solution to Activity 11

The data are measurements of the speed of light in air, and are therefore best represented by a continuous variable. Thus the model should be a continuous one. Michelson's measurements might be expected to cluster around some average value, close to the true speed of light in air. There is no particular reason to suppose that the measurements are more likely to fall below or above this average, and hence a symmetric distribution seems reasonable. Moreover, the data are sample means of other values. Thus these data appear to conform closely to the standard setting for the normal distribution, given in Subsection 3.2 of Unit 12. A possible choice of model is therefore the normal distribution. (Because the measurements will take enormous values, any reasonable normal model will give infinitesimal probability to negative values.)

Solution to Activity 12

Graph > Histogram...

(a) A histogram of the data is shown in Figure 19.

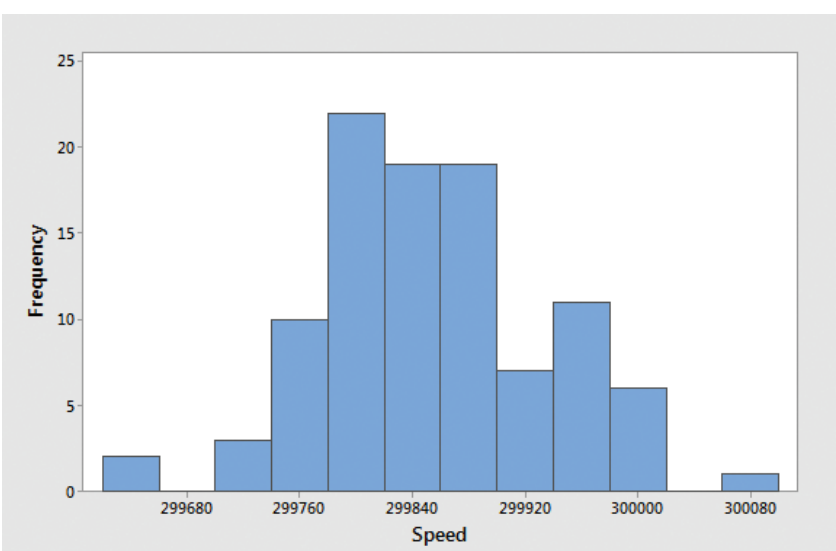


Figure 19 A histogram of the speed of light data

The histogram is broadly symmetric and unimodal. In these respects, the normal model seems appropriate.

(b) The validity of the normal model can be checked by means of a normal probability plot. Such a plot is shown in Figure 20.

Graph > Probability Plot...

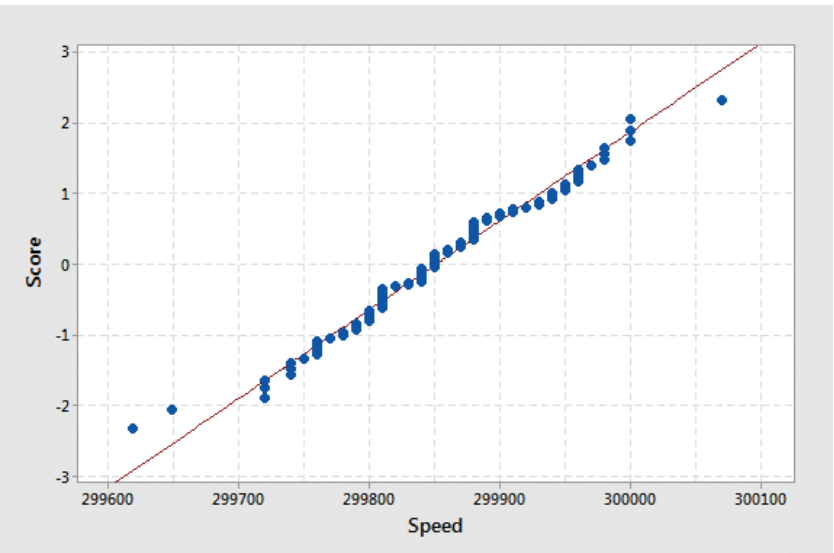


Figure 20 A normal probability plot of the speed of light data

The points on this plot seem to be in a reasonably straight line, except only perhaps for the most extreme values. Based on this plot, too, the normal model seems appropriate.

Solution to Activity 13

(a) The sample mean of Michelson's measurements of the speed of light in air is 299 851 $\mathrm{km\,s^{-1}}$.

Stat > Basic Statistics > Display Descriptive Statistics...

If you assume a normal model, then you can calculate a t-interval for the mean. The 95% confidence interval is (299 836, 299 867), and the 99% confidence interval is (299 831, 299 872).

Use **Stat > Basic Statistics > 1-Sample t...** (see, for example, Activity 21 of Computer Book B).

(b) The true value of the speed of light in air is around 299 703 $\mathrm{km\,s^{-1}}$. This value lies outside the 99% confidence interval (and hence outside the 95% confidence interval also). Nevertheless, the discrepancy between Michelson's estimate and the true value is only $299\,851 - 299\,703 = 148$ $\mathrm{km\,s^{-1}}$. The error, expressed as a proportion of the true value, is $148/299\,703 \simeq 0.00049$, or 0.049%. (This is called the *relative error*.) In this sense, Michelson's measurement was remarkably accurate, even though it is not accurate enough in the sense that the true value is not contained within the 99% confidence interval.

(c) The appropriate test is a two-sided (one-sample) t-test. Minitab gives 0.000 for the p-value, so $p < 0.0005$. There is strong evidence that the underlying mean of Michelson's measurements is not equal to the true speed of light in air.

Use **Stat > Basic Statistics > 1-Sample t...**, enter `299703` (the true value) in the **Hypothesized mean** field, and set **Alternative hypothesis** to `Mean ≠ hypothesized mean`.

(d) Michelson's experiment was very accurate given the methods available to him, in that he succeeded in measuring the speed of light with a relative error of less than 0.05% (see the solution to part (b)). However, the discrepancy is unlikely to be due to random measurement error, since the true value is not contained in the 99% confidence interval, and a hypothesis test provided strong evidence against the null hypothesis of equality. Thus it is likely that small but systematic biases occurred in Michelson's experiment, producing a slight overestimate in the true value of the speed of light in air.

Solution to Activity 14

Graph > Scatterplot...

(a) A scatterplot is a suitable graphical display for this purpose. The scatterplot of speed measurements against day of experiment is shown in Figure 21.

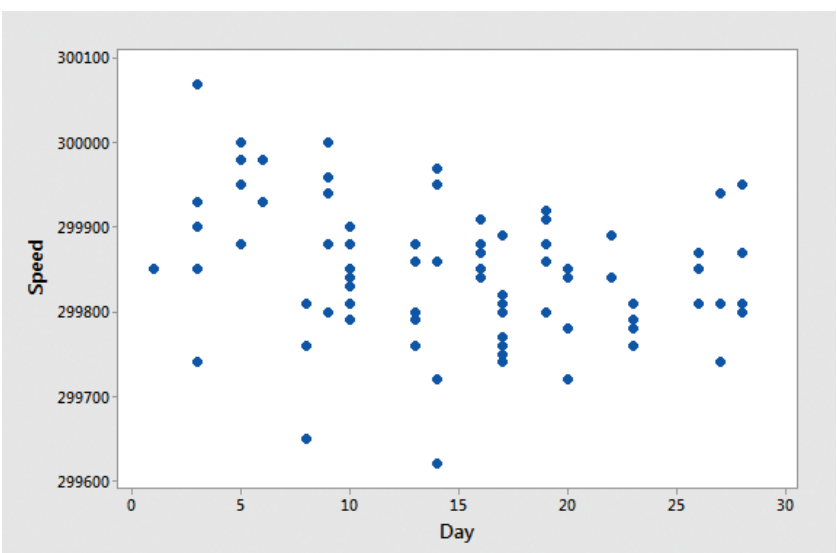


Figure 21 Speed measurement against day of experiment

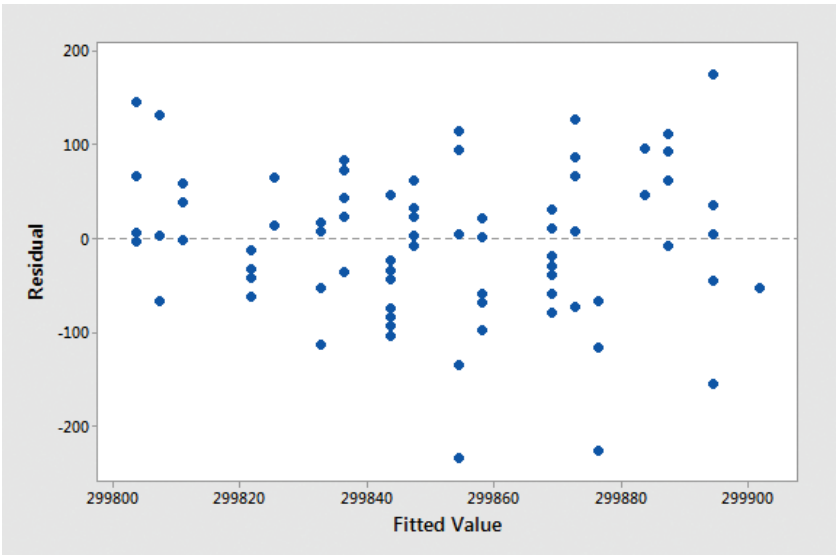
It is hard to be sure what to make of this scatterplot. Do you perceive a trend, or is there none? Arguably, the scatterplot shows a general downward tendency overall. If so, it would appear that, on average, Michelson's measurements tended to decrease over time. (But, then, is there a flattening out, or even a slight increase, towards the end?)

Stat > Regression > Regression > Fit Regression Model... The plots arise from choosing options in the **Regression: Graphs** dialogue box.

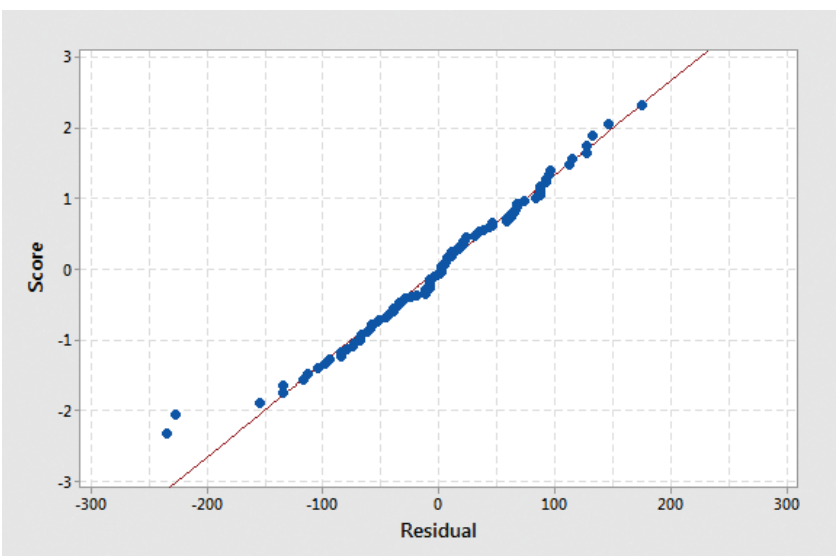
(b) The least squares line fitted to these data is

$$\text{Speed} = 299\,906 - 3.64\,\text{Day}.$$

A residual plot and a normal probability plot of the residuals are shown in Figure 22.



(a)



(b)

Figure 22 (a) Residual plot (b) normal probability plot of residuals

The points in the residual plot appear to be scattered reasonably randomly about zero, suggesting that the assumption that the random terms have constant, zero mean and constant variance seems plausible. (On closer inspection, the possible proviso of a slight broadening of the band of residuals as the fitted values increase seems to hint more at a 'high frequency' variation in the trend, and hence the mean, rather than a slight increase in the variance; any such effect seems small and not important to worry about in the current context.)

With only a couple of exceptions, the residuals lie close to a straight line in the normal probability plot, so the assumption that the random terms are normally distributed seems plausible.

(c) The p-value obtained from Minitab is $p = 0.001$, which is much less than 0.01 and hence corresponds to strong evidence against the null hypothesis of no linear regression relationship. Since the slope of the least squares line is negative, it seems that Michelson's measurements of the speed of light decreased in size as his series of experiments progressed.

(d) According to the model, for each additional day, the speed measurements decreased, on average, by 3.64 $\mathrm{km\,s^{-1}}$.

Solution to Activity 15

(a) A scatterplot of speed against temperature is shown in Figure 23.

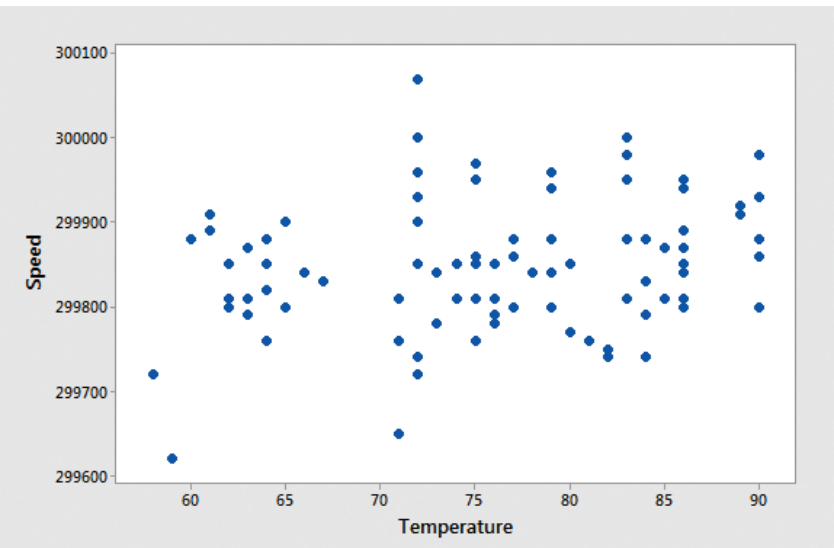


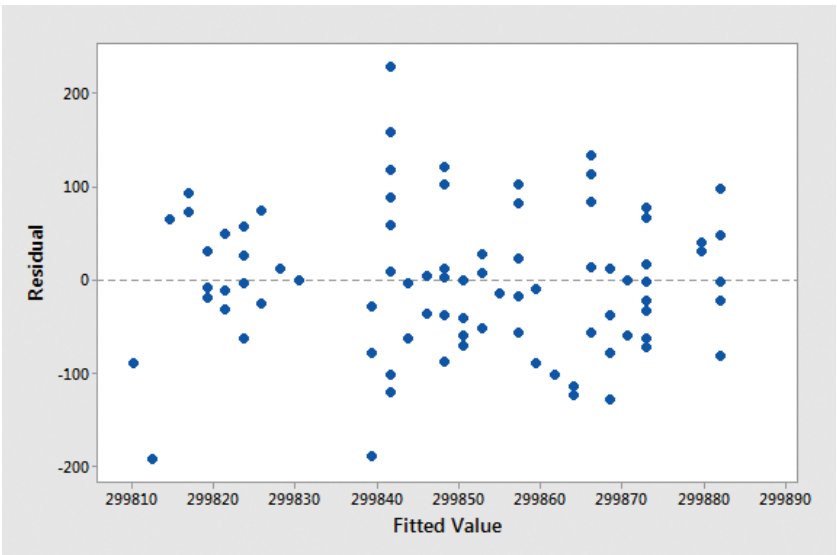
Figure 23 Speed measurement against temperature

It is not obvious that there is any relationship between speed measurement and ambient temperature. The plot perhaps suggests a slightly increasing trend. One or two outliers are perhaps suggested, also.

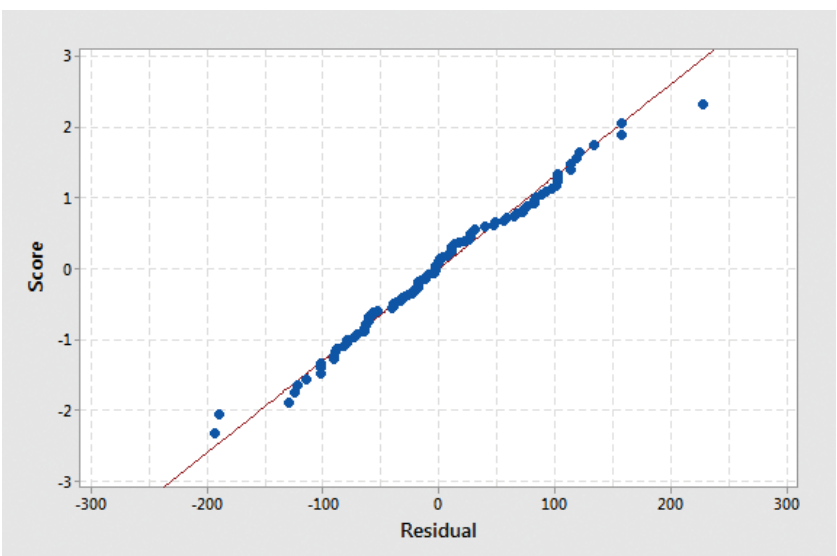
(b) The least squares line fitted to these data is

$$\text{Speed} = 299\,680 + 2.242\,\text{Temperature}.$$

A residual plot and a normal probability plot of the residuals are shown in Figure 24.



(a)



(b)

Figure 24 (a) Residual plot (b) normal probability plot of residuals

In general, the points in the residual plot appear to be scattered reasonably randomly about zero, and the residuals lie close to a straight line in the normal probability plot, so the assumptions concerning the random terms seem plausible. The only possible deviation from these claims remains the very few potential outliers.

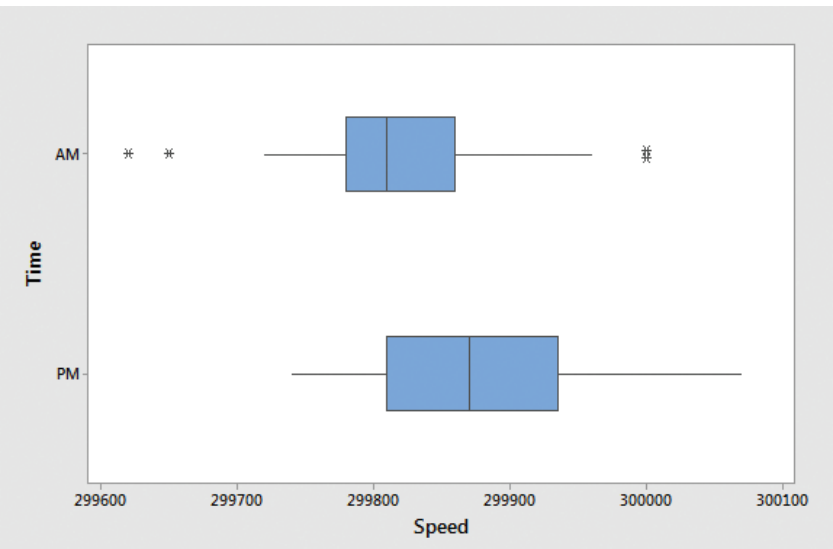
(c) The p-value obtained from Minitab is $p = 0.014$. Since $0.01 < p < 0.05$, there is moderate evidence against the null hypothesis of no regression relationship. There is some evidence that, despite his best efforts, Michelson's measurements of the speed of light did indeed increase in size when temperatures were higher.

(d) According to the model, for each degree rise in temperature, the speed measurements increased, on average, by 2.242 $\mathrm{km\,s^{-1}}$.

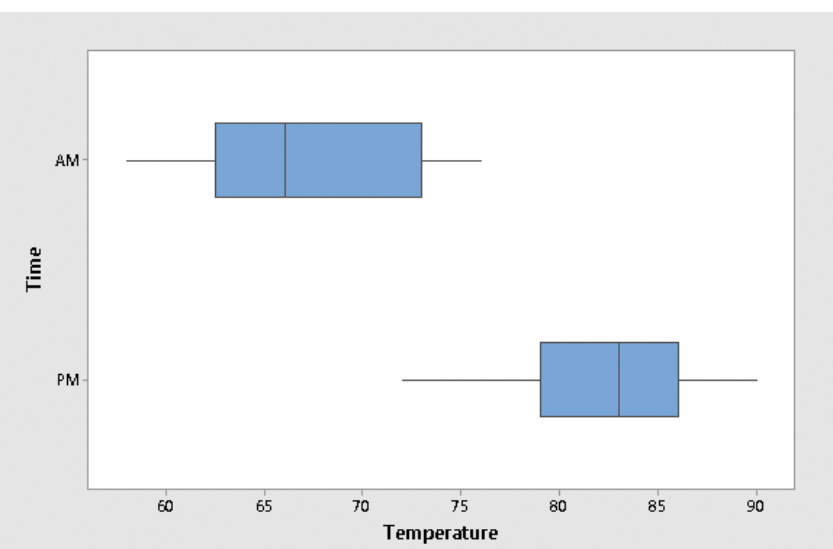
Solution to Activity 16

(a) Clicking on the **Columns** subfolder of the **michelson.mtw** folder in the Project Manager provides a summary containing information including missing values. There are just two missing `Time` values.

(b) The comparative boxplots are shown in Figure 25.



(a)

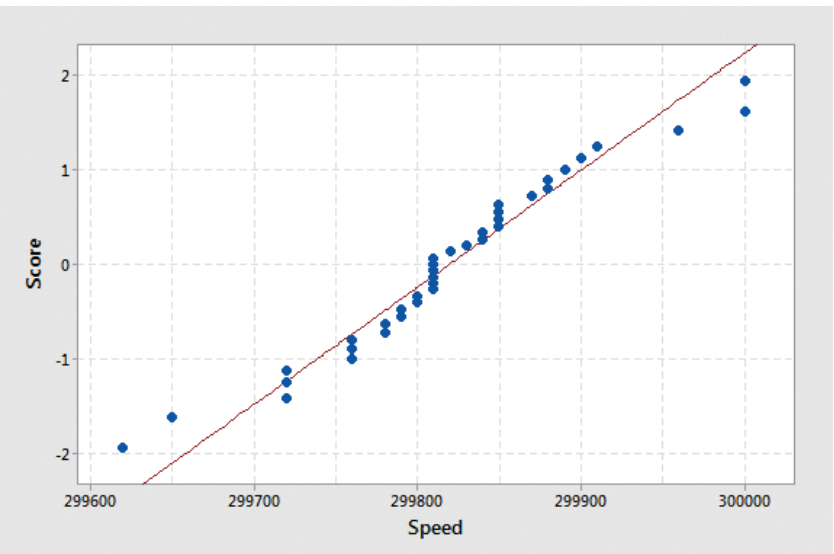


(b)

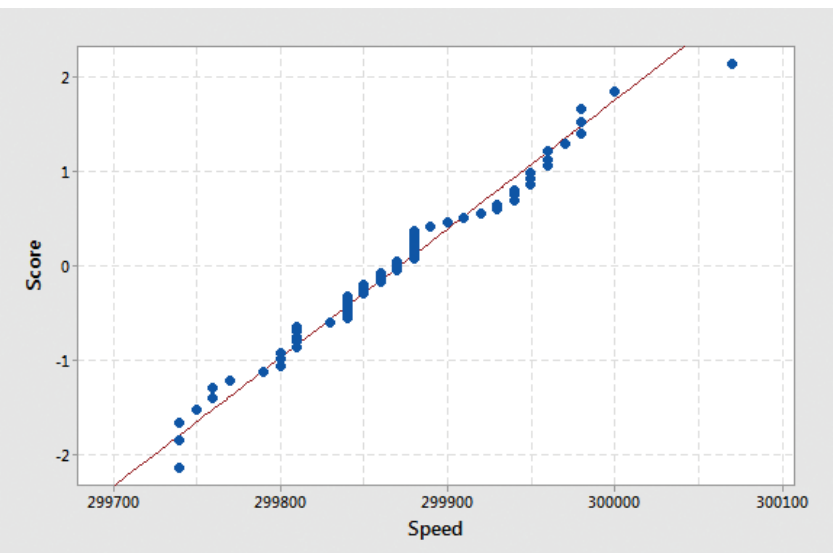
Figure 25 Comparative boxplots of (a) speed measurements (b) temperature

The boxplots suggest that the speed measurements are a little higher when measured in the (late) afternoon than in the (early) morning. There are also some possible outliers in the morning group. As might be expected, the ambient temperature is higher in the late afternoon than in the early morning.

(c) The normal probability plots are shown in Figure 26.



(a)



(b)

Figure 26 Normal probability plots of speed measurements in (a) the morning (b) the afternoon

The plots broadly suggest that the normality assumption appears to be reasonable for both groups. The measurements in the morning are perhaps 'a little less normal' than those in the afternoon, especially with the possibility observed in part (b) of some outliers.

Stat > **Basic Statistics** > **Display Descriptive Statistics**..., entering `Time` in the **By variables (optional)** field.

(d) The variances of the speed measurements in the two groups are 6503 (`AM`) and 5359 (`PM`). The ratio of the larger variance to the smaller is $6503/5359 \simeq 1.21$. This is less than 3 so, by the rule of thumb given in Subsection 4.4 of Unit 8, the assumption of equal variances seems reasonable.

(e) The required confidence interval is $(-83.2, -20.3)$ km s^{-1}. Because zero is not in the confidence interval, it suggests that there is a real difference between the speed measurements according to time of day, with measurements taken in the late afternoon tending to be higher than those taken in the morning.

(f) It seems that there is a difference between speed measurements obtained at different times of day, and that this difference may be related, at least in part, to temperature effects.

Solutions to exercises

Solution to Exercise 1

(a) Choose **Stat** > **Regression** > **Regression** > **Fit Regression Model...** with `Coaching` in the **Responses** field and `Membership` in the **Continuous predictors** field. The least squares line for the data is given in the Minitab Session window as

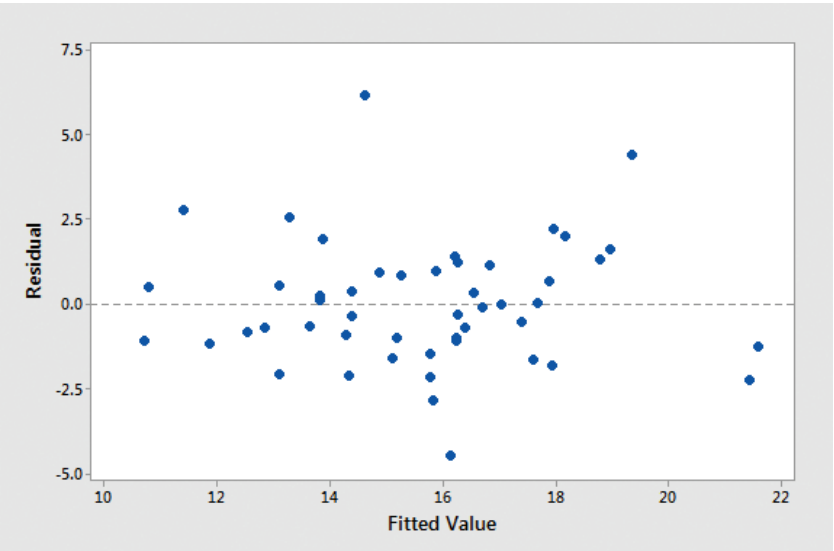
$$\text{Coaching} = -5.82 + 0.989\,\text{Membership}.$$

(b) The value $\widehat{\alpha} = -5.82$ is the estimated value of the intercept; it is of no interest here because it corresponds to sports partnership areas with no sports club members.

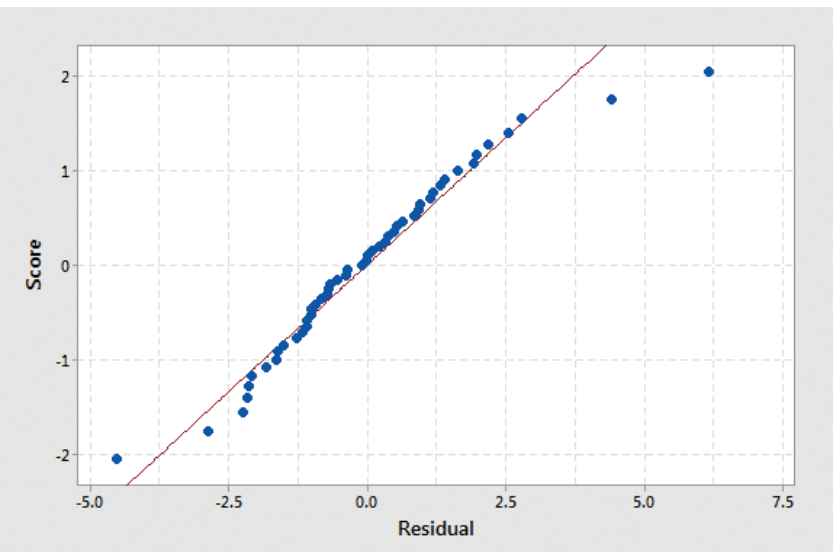
The value $\widehat{\beta} = 0.989$ is the estimated value of the slope. It estimates that, for each 1% increase in sports club membership in a sports partnership area, one might expect to see a 0.989% increase in the percentage of adults who received coaching. Notice that, in this case, the slope is almost one. So you could suggest that for each 1% increase in sports club membership in a sports partnership area, one might expect to see a corresponding 1% increase in the percentage of adults who received coaching.

(c) When the options **Normal probability plot of residuals** and **Residuals versus fits** are chosen in the **Regression: Graphs** dialogue box, a residual plot and a normal probability plot for the residuals are obtained. They are shown in Figure 27 (overleaf).

There is no obvious pattern in the residual plot in Figure 27(a), so the assumption of constant, zero mean and constant variance of the random terms seems appropriate. The points in the normal probability plot in Figure 27(b) lie roughly along a straight line, so the assumption of normality of the random terms seems plausible. The only departures correspond to individual points which might, as the question suggested, be considered to be outliers. In its Session window output, Minitab drew attention to three points with large residuals (one negative, the leftmost on the probability plot, and two positive, the rightmost two on the probability plot) as well as two other points it considered to be unusual due to their large fitted values (but their residuals are not large – they are the rightmost points on the residual plot). Apart possibly from one or two of the points identified with large residuals, the assumptions underlying the linear regression model seem to hold quite well for these data.



(a)



(b)

Figure 27 (a) Residual plot (b) normal probability plot of residuals

Solution to Exercise 2

(a) Choose **Stat** > **Regression** > **Regression** > **Fit Regression Model...** with `Price` in the **Response** field, and `Age` and `Bidders` in the **Continuous predictors** field. In the **Regression: Graphs** dialogue box, select **Normal probability plot of residuals** and **Residuals versus fits**. From the Minitab output, the fitted multiple regression model is

$$\text{Price} = -1337 + 12.736\,\text{Age} + 85.82\,\text{Bidders}.$$

(b) The output table labelled '`Coefficients`' is reproduced below.

```
Coefficients
Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant   -1337      173    -7.71    0.000
Age        12.736    0.902    14.11    0.000  1.07
Bidders     85.82     8.71     9.86    0.000  1.07
```

From this, the p-values associated with both regression coefficients are given as 0.000. There is therefore strong evidence to suggest that both regression coefficients are non-zero.

(c) The interpretation of the regression coefficients is as follows.

- If the age of the clock (x_1) increases by one year, and the number of bidders (x_2) remains fixed, then the selling price of the clock (y) would be expected to increase by £12.736.
- If the number of bidders (x_2) increases by one, and the age of the clock (x_1) remains fixed, then the selling price of the clock (y) would be expected to increase by £85.82.

(d) A clock with $x_1 = 150$ and $x_2 = 6$ is predicted to have a sale price, in pounds, of

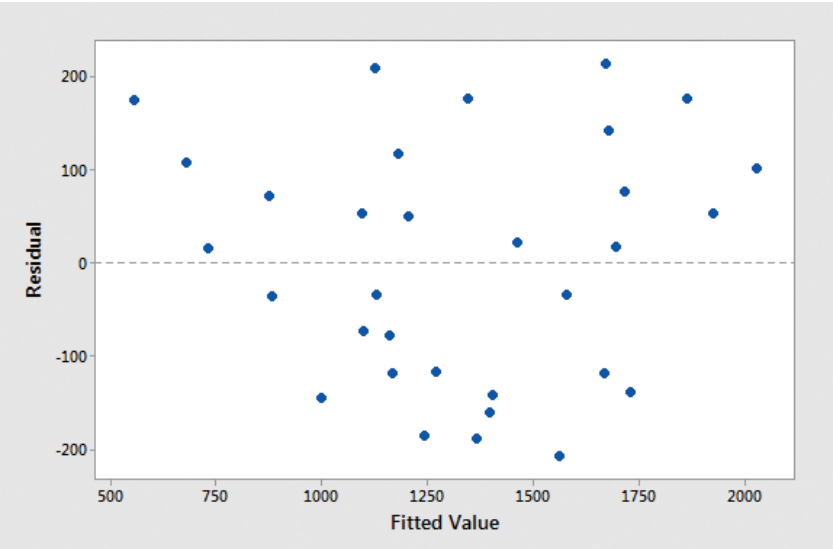
$$y = -1337 + 12.736 \times 150 + 85.82 \times 6 = 1088.32.$$

(e) No unusual points are highlighted in the Minitab output in the Session window.

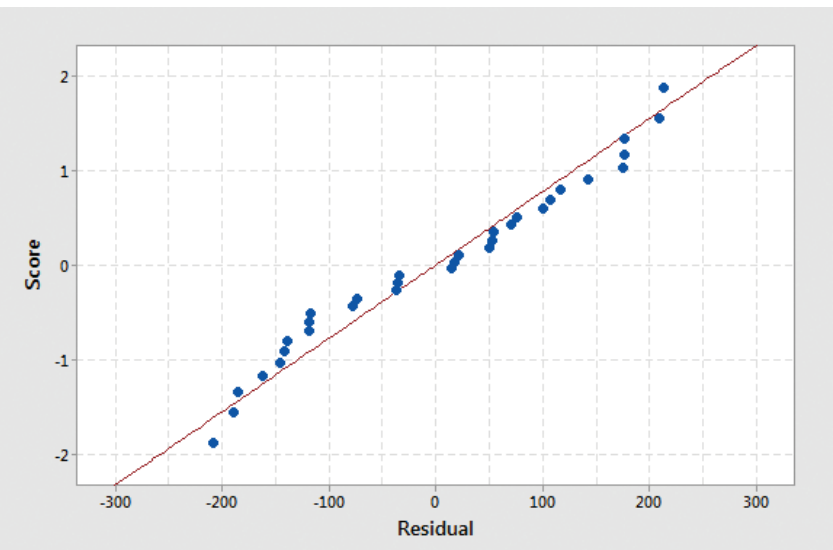
The residual plot and normal probability plot of residuals are shown in Figure 28 (overleaf).

The points in the residual plot seem to be fairly randomly scattered about zero, and so the assumption that the random terms have constant, zero mean and constant variance also seems reasonable.

The points in the normal probability plot track out something of an 'S-shape' rather than following the straight line that one would expect if the assumption that the random terms are normally distributed is good. It seems that possibly a non-normal distribution might be more appropriate for the random terms, although any departure from normality may not be sufficiently extreme for an assumption of normality to give misleading inferences.



(a)



(b)

Figure 28 (a) Residual plot (b) normal probability plot of residuals

Solution to Exercise 3

(a) Using **Graph** > **Scatterplot. . .**, a scatterplot of cranial capacity against age can be obtained and is shown in Figure 29.

The data appear to follow something of a curve rather than a straight line, but with a reasonably constant variation about that curve. The cranial capacity of the human skull decreases with age of the skull, more quickly for lower ages, more slowly for higher ages.

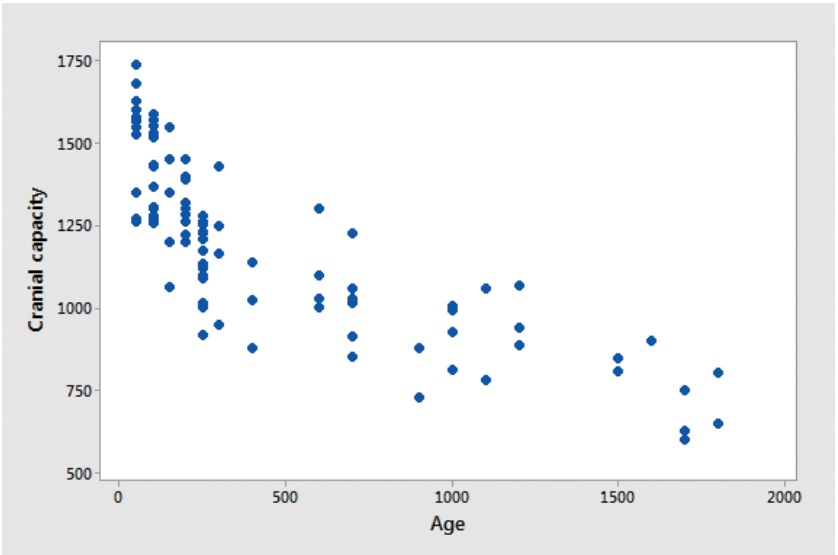


Figure 29 Cranial capacity against age

(b) (i) Use **Calc** > **Calculator...** to obtain the transformed versions of the explanatory variable, `Age`. The three scatterplots will be as shown in Figures 30, 31 and 32.

In Figures 30 and 31 (overleaf), cranial capacity is plotted against the square root of age and the log of age, respectively. In both plots the points lie fairly close to a straight line. It seems that both a square root and a logarithmic transformation might be appropriate for straightening out the non-linear relationship in the original data.

Use `Natural log (log base e)` to obtain log(age).

Figure 30 Cranial capacity against $\sqrt{\text{age}}$

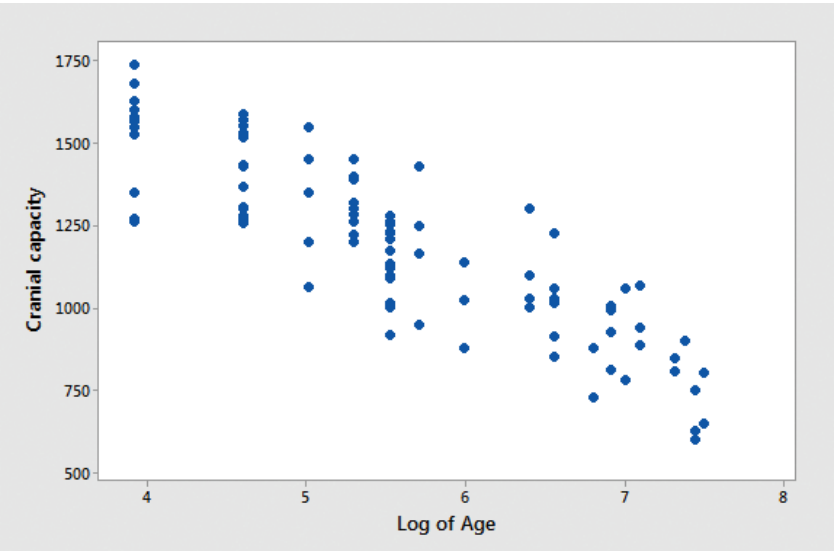


Figure 31 Cranial capacity against log(age)

In Figure 32, cranial capacity is plotted against the reciprocal of age. The points seem to follow a curve rather than a straight line. So a reciprocal transformation of the explanatory variable is not appropriate.

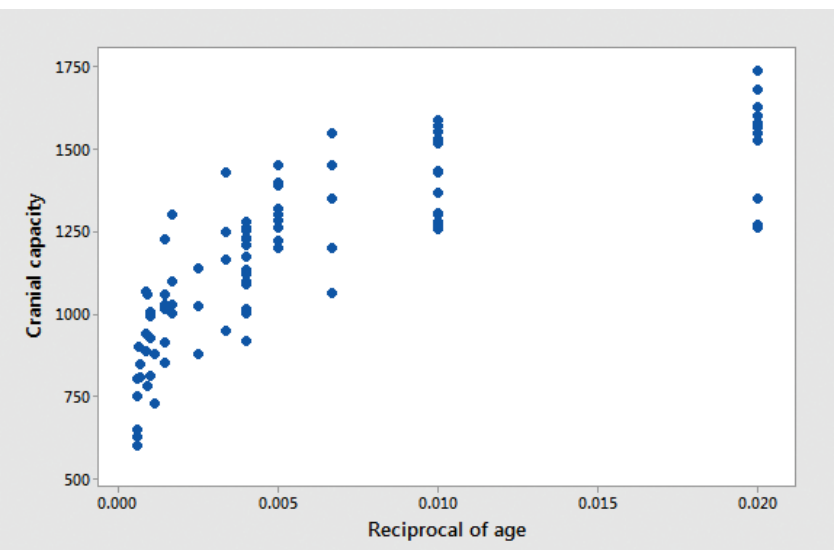


Figure 32 Cranial capacity against 1/age

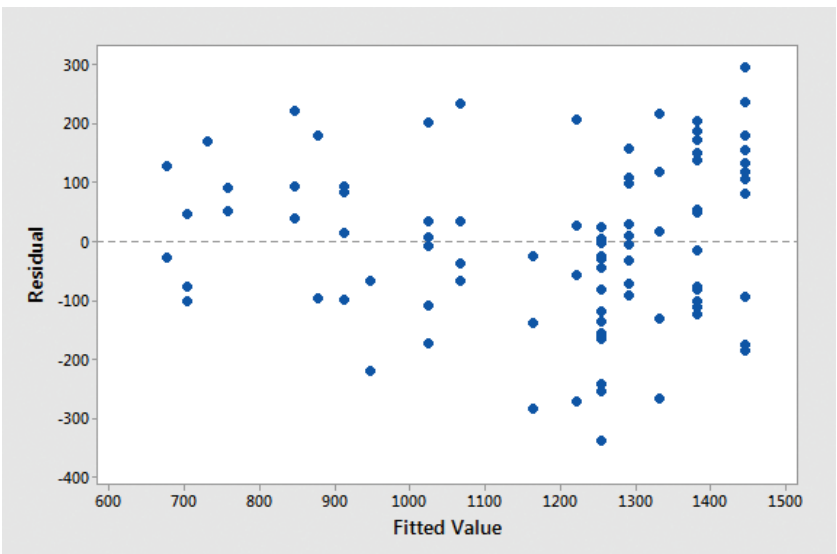
(ii) Both the first and second transformations seem to straighten out the non-linear relationship reasonably well, so either could reasonably be chosen, whereas the reciprocal transformation does not seem to be appropriate.

(iii) Use **Stat** > **Regression** > **Regression** > **Fit Regression Model...**, obtaining the required graphs by clicking on **Graphs...** to obtain the **Regression: Graphs** dialogue box.

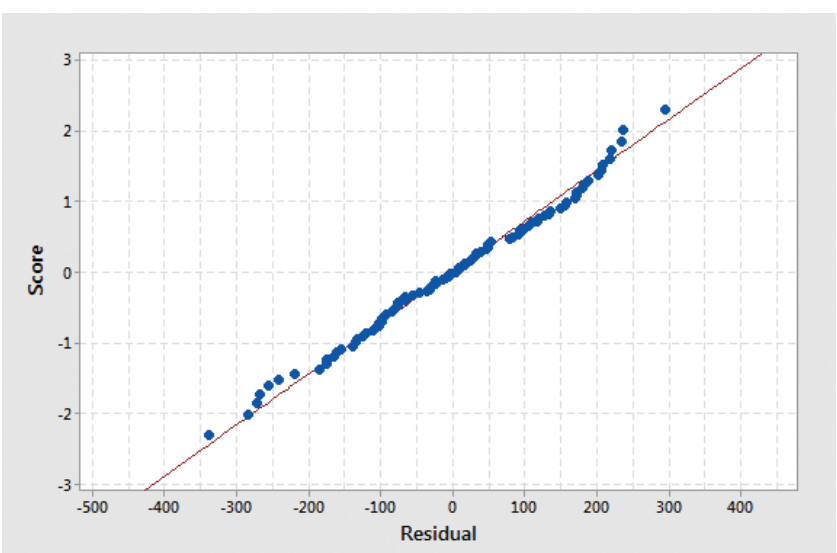
If you opted for the first transformation, the square root, then the equation of the least squares line for the transformed data would be

$$\text{Cranial capacity} = 1598.5 - 21.73\sqrt{\text{Age}}.$$

A residual plot and a normal probability plot of the residuals are shown in Figure 33.



(a)



(b)

Figure 33 (a) Residual plot (b) normal probability plot; square root transformation

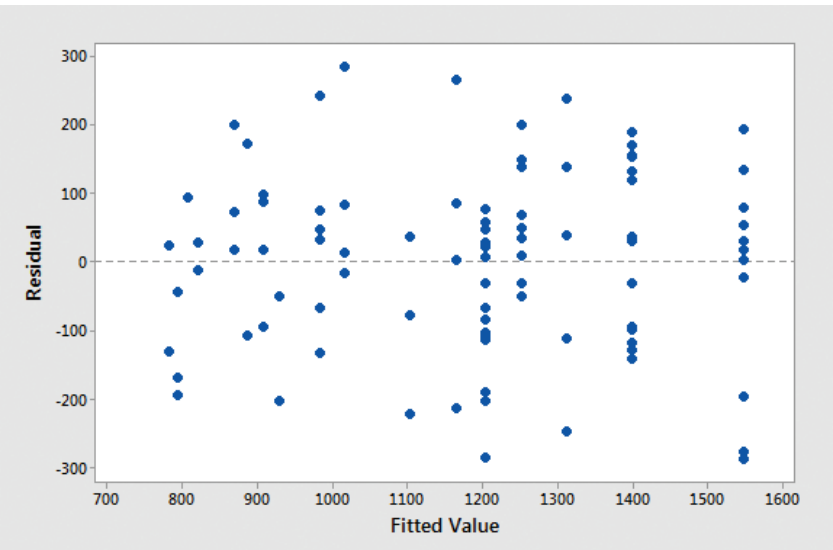
You might, or might not, perceive a slightly curved pattern in the residual plot. But the assumption of constant, zero mean and constant variance of the random terms appears to be plausible.

(If you perceive an increase in variance towards the right of the residual plot, you might be right, or it might be a false impression caused by there being more datapoints in this region.) The normal probability plot clearly shows the assumption of normality of the random terms to be plausible.

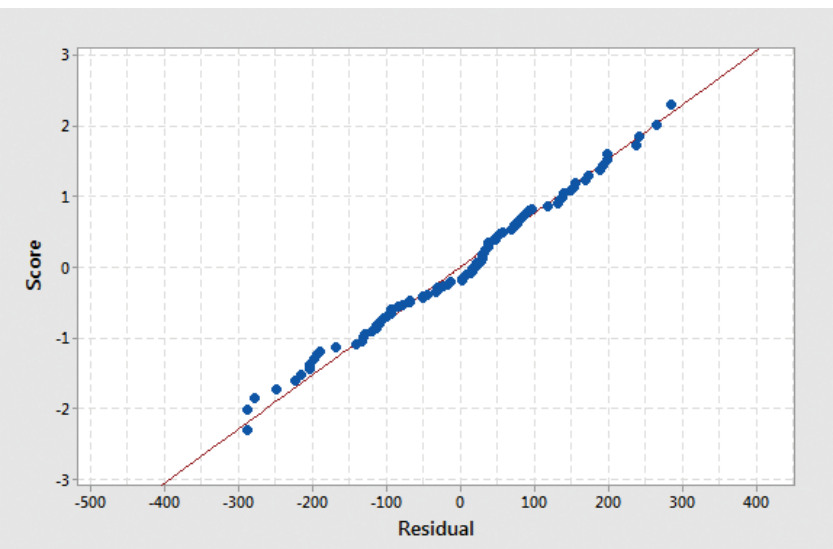
If you opted for the second transformation, log, then the equation of the least squares line for the transformed data would be

$$\text{Cranial capacity} = 2383.3 - 213.8 \log(\text{Age}).$$

A residual plot and a normal probability plot of the residuals are shown in Figure 34.



(a)



(b)

Figure 34 (a) Residual plot (b) normal probability plot; log transformation

The points in the residual plot appear to be scattered randomly so the assumption of constant, zero mean and constant variance of the random terms seems appropriate. The normal probability plot also shows the assumption of normality of the random terms to be plausible.

Either a square root or a log transformation is suitable in this case.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Page 5: © Trickyboy This file is licensed under the Creative Commons Attribution-ShareAlike Licence
http://creativecommons.org/licenses/by-sa/3.0/

Page 11: © Stuart Milligan / Getty Images

Page 13: © Ben Holland This file is licensed under the Creative Commons Attribution-ShareAlike Licence
http://creativecommons.org/licenses/by-sa/4.0/

Page 15: © Bizoon / www.123rf.com

Page 17: © Luchschen / www.123rf.com

Page 19: © Science History Images / Alamy Stock Photo

Page 21: Taken from:
http://www.rpi.edu/dept/phys/Dept2/APPhys1/optics/optics/node4.html

Page 25: © chas73 / www.123rf.com

Page 26: Lee, S.-H. and Wolpoff, M.H. (2003) 'The pattern of evolution in Pleistocene human brain size', *Paleobiology*, vol. 29, no. 2, pp. 186–96

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked, the publishers will be pleased to make the necessary arrangements at the first opportunity.

Index